Polygenic risk modeling with latent trait-related genetic components

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

Author affiliations

[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]Department of Statistics, Stanford University, Stanford, CA, 94305, USA.

Abstract:

Polygenic risk models have led to significant advances in understanding complex diseases and their clinical presentation. While traditional models of genetic risk like polygenic risk scores (PRS) 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 Decomposition of Genetic Associations (DeGAs), which we call the DeGAs polygenic risk score (dPRS). We compute DeGAs on associations from 1,905 traits in the UK Biobank and find that dPRS performs comparably to standard PRS while offering greater interpretability. We highlight results for body mass index (BMI), myocardial infarction (heart attack), and gout in 337,151 white British individuals (spilt 70/10/20 for training, validation, and testing), with replication in a further set of 25,486 non-British whites from the Biobank. We show how to decompose an individual's genetic risk for a trait across these latent components. For example, we find that BMI polygenic risk factorizes into distinct components relating to fat-free mass, fat mass, and overall health indicators like sleep duration and alcohol and water intake. Most individuals with high dPRS for BMI have strong contributions from both a fat mass component and a fat-free mass component, whereas a few 'outlier' individuals have strong contributions from only one of the two components. Our methods enable fine-scale interpretation of the drivers of genetic risk for complex traits.

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

Introduction:

Common diseases 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 toward phenotypes of interest, 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]. Although many versions of PRS can be used to estimate risk[11–13], previous work suggests 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 a trait as a sum of effects from contributing genetic variants, the DeGAs polygenic risk score (dPRS) models genetic risk for traits as a sum of contributions from DeGAs components[15], each consisting of a set of variants with consistent effects on a subset of the traits being modeled (**Figure 1**). Genetic risk for an individual DeGAs component can be expressed as a component PRS (cPRS) that approximates risk for a weighted combination of relevant traits. We then use these scores to estimate personalized genetic risk profiles that inform genetic subtyping of disease.

As proof of concept, we compute DeGAs[15] using summary statistics generated from genome-wide associations between 1,905 traits and 454,565 independent common variants in a subset of unrelated white British individuals (*n*=236,005) in the UK Biobank[16] (**Methods**). We then develop a series of dPRS models and evaluate their performance in independent samples of unrelated individuals in the same population (*n*=33,716 validation set; *n*=67,430 test set), and in UK Biobank non-British whites (*n*=25,486 additional test set). We highlight results for body mass index (BMI), myocardial infarction (MI/heart attack), and gout, motivated by their high prevalence (obesity, in the case of BMI) among older individuals in this cohort[17].

Results:

Evaluating the DeGAs Polygenic Risk Score (dPRS):

Genome-wide associations between *n*=1,905 traits and *m*=454,565 independent human leukocyte antigen (HLA) allelotypes, copy-number variants[18], and array-genotyped variants were computed in a population of 236,005 unrelated white British individuals from the UK Biobank study[16] (**Methods**). We applied DeGAs[15] to beta- or *z*-statistics from these GWAS with varying *p*-value thresholds for input (**Figure 1a**). We then defined polygenic risk scores for each DeGAs component (cPRS, **Figure 1b**) and used them to build the DeGAs polygenic risk score (dPRS; **Figure 1c**). The model with optimal out of sample prediction (**Supplementary Figure 1**) corresponded to DeGAs on *z*-statistics with nominally significant ($p < 0.01$) associations.

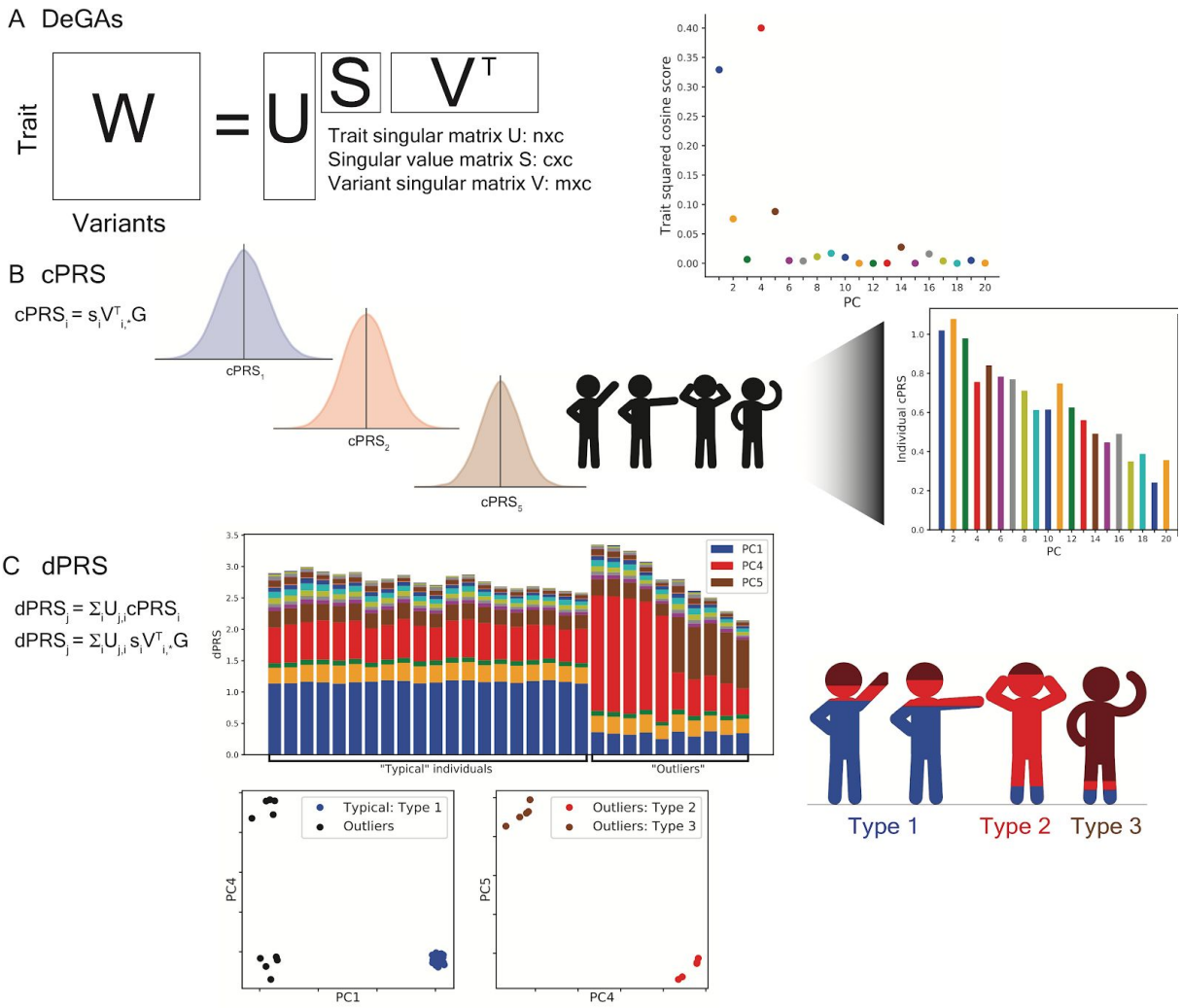

Figure 1: Study overview. (A) Matrix Decomposition of Genetic Associations (DeGAs) is performed by taking the truncated singular value decomposition (TSVD) of a matrix *W* containing summary statistics from GWAS of *n*=1,905 traits over *m*=454,565 variants from the UK Biobank. The squared columns of the resulting singular matrices *U* (*n x c*) and *V* (*m x c*) measure the importance of traits (variants) to each component; the rows map traits (variants) back to components. The squared cosine score (a unit-normalized row of *US*) for some hypothetical trait indicates high contribution from PC1, PC4, and PC5. **(B) Component polygenic risk scores (cPRS)** for the *i*th component is defined as $s_i V^T_{i,*} G$ (*i*-th singular value in *S* and *i*-th row in V^T), for an individual with genotypes *G*. **(C) DeGAs polygenic risk scores (dPRS)** for trait *j* are recovered by taking a weighted sum of $cPRS_i$, with weights from *U* (*j,i*-th entry). We also compute DeGAs risk profiles for each individual (**Methods**), which measure the relative contribution of each component to genetic risk. We "paint" the dPRS high risk individuals with these profiles and label them "typical" or "outliers" based on similarity to the mean risk profile (driven by PC1, in blue). Outliers are clustered on their profiles to find additional genetic subtypes: this identifies "Type 2" and "Type 3", with risk driven by PC4 (red) and PC5 (tan). Clusters visually separate each subtype along relevant cPRS (below). Image credit: VectorStock.com/1143365.

To validate this model, we estimated disease prevalence (or, for BMI, mean BMI) at several quantiles of risk in a held-out test set of white British individuals in the UK Biobank (*n*=67,430). For all traits we observed increasing severity (quantitative) or prevalence (binary) at increasing quantiles of dPRS (**Figure 2a-c**) adjusted for age, sex, and the first 4 genotype principal components from UK Biobank's PCA calculation[16]. This trend was most pronounced at the highest risk quantile (2%) for each trait. At this stratum we observed 2.47 kg/m^2 higher BMI (95% CI: 2.21-2.74); 1.47-fold increased odds of MI (CI: 1.14-1.89); and 3.10-fold increased odds of gout (CI: 2.35-4.09) over the population average in the test set.

Further, we found dPRS to be comparable to prune- and threshold-based PRS using the same input data (**Supplementary Figure 2**). Although there was some discrepancy between the individuals considered high risk by each model (**Supplementary Figure 3, Supplementary Table 1**) we observe similar effects at the extreme tail of PRS as with dPRS. The top 2% of risk of PRS for each trait had 2.48 kg/m^2 higher BMI; 1.63-fold increased odds of MI (CI: 1.27-2.08); and 2.22-fold odds of gout (CI: 1.64-2.99) (**Figure 2a-c**) using the same covariate adjustment as dPRS. Population-wide predictive measures were also similar, with BMI residual *r*=0.205, and PRS AUC (not adjusted for covariates) 0.561 for MI and 0.605 for gout (**Figure 2d-f**). 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 traits, and note a similar trend for the other DeGAs traits (**Supplementary Figure 2, Supplementary Data 1**).

However, we note that dPRS and PRS add comparatively little population-wide predictive value over factors such as age, sex, and demographic effects that are captured by genomic PCs (**Figure 2d-f**). At the population level, we found *r*=0.207 between covariate-adjusted dPRS and residual BMI. For binary traits we estimated an area-under receiver operating curve (AUC) of 0.560 for MI and 0.645 for gout, using unadjusted dPRS as the classifying score. After adjustment for covariates, the marginal increase in AUC is modest: 0.005 for MI and 0.035 for gout.

Characterizing DeGAs Components:

We describe the latent structure identified through DeGAs by annotating each component with its contributing traits (**Figure 3**) and variants, aggregated by gene (**Supplementary Data 2,3**). The relative importance of traits to components is measured using the trait contribution score[15], which corresponds to a squared column of the trait singular matrix *U*. The relative importance of components to each trait is measured using the trait squared cosine score[15], which is taken from a normalized squared row of *US*. These scores are defined for variants and genes using the variant singular matrix *V*.

Body mass index is a highly polygenic trait with associated genetic variation relevant to adipogenesis, insulin secretion, energy metabolism, and synaptic function[15,19]. Here, the DeGAs trait squared cosine score (**Figure 3a**) indicates strong contribution from components related to

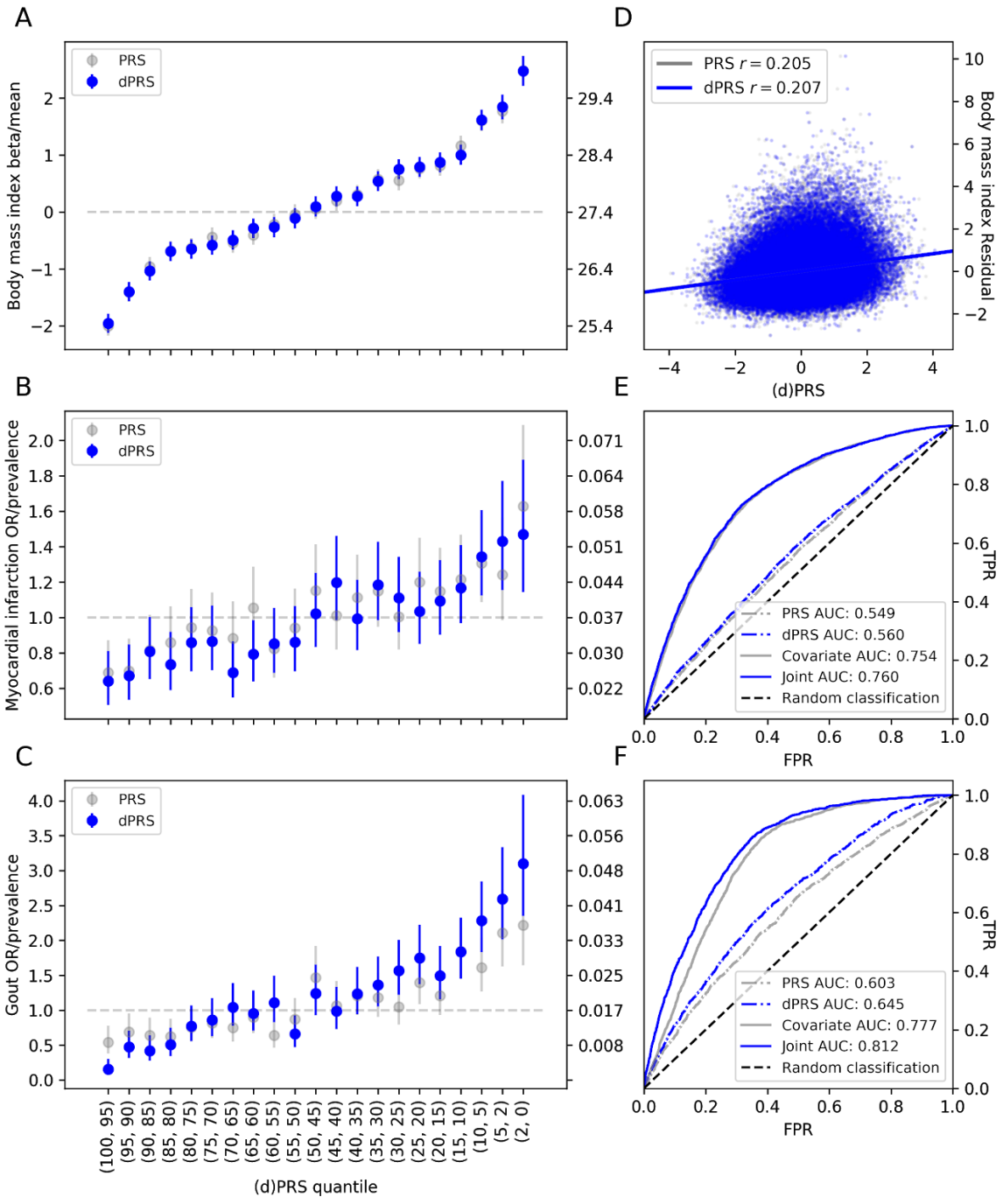

Figure 2: Performance of dPRS. (A-C) Effect of increased risk (dPRS or PRS) on BMI, MI, and gout. Beta/OR (left axis) were estimated by comparing the quantile of interest (x-axis) with a middle quantile (40-60%), adjusted for these covariates: age, sex, 4PCs (**Methods**). Trait mean or prevalence (right axis) was computed within each quantile; error bars denote the 95% confidence interval of each estimate. **(D)** Correlation between dPRS or PRS and covariate adjusted BMI. Receiver operating curves with area under curve (AUC) values for MI **(E)** or gout **(F)** for dPRS, PRS, covariates, and a joint model with covariates and dPRS. Models with covariates were fit in the validation set; all evaluation was in the test set. (**Methods**).

body size and fat-free mass (PC1 - 42.8%), fat mass (PC3 - 37.9%), and overall health indicators like sleep duration and alcohol and water intake (PC169 - 1.8%). Genic variation proximal to *FTO* and *DLEU1* had the highest contribution to PC1 (**Supplementary Data 3**). Variants proximal to both genes are strongly associated with traits affecting body size in adults[20,21]. The former is an alpha-ketoglutarate dependent dioxygenase whose causal role in BMI has been questioned[22]; the latter is a tumor-suppressing lncRNA named for its frequent deletion in patients with chronic lymphocytic leukemia[23]. Genetic variants proximal to *DLEU1* are also significant contributors to PC3 along with *TSBP1* (or *C6orf10*), an open reading frame in the human leukocyte antigen (HLA) region.

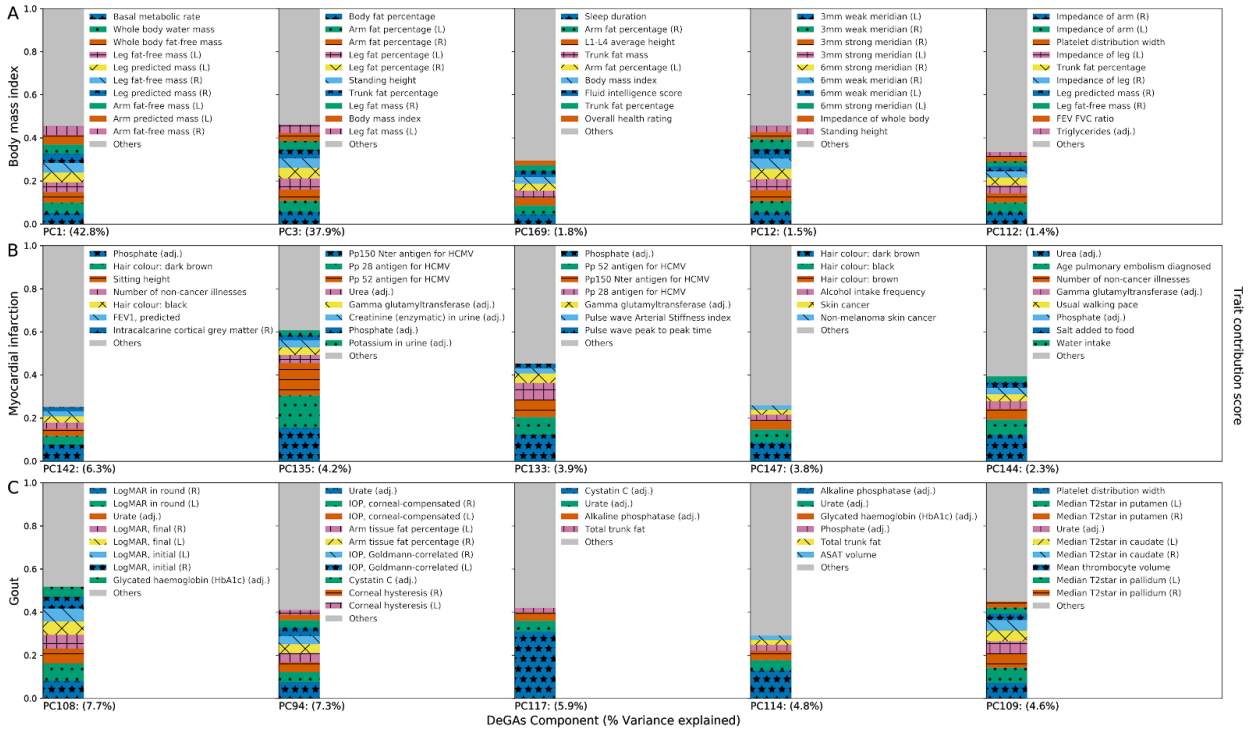

Figure 3: Top five DeGAs components for each trait. (A) Top five DeGAs components for body mass index, as ranked by the trait squared cosine score. Each component is labeled with its top ten traits, as determined by the contribution score (squared column of *U*), and with the fraction of variance in genetic associations it explains (squared cosine score). Traits are displayed for a component if their contribution score for the component exceeds 0.02. Top five components for **(B)** myocardial infarction and **(C)** gout.

Myocardial infarction is similarly a polygenic outcome with well-established risk factors attributable to common and rare genetic variation[8], age, sex, and lifestyle attributes like diet and smoking. DeGAs components important to this trait are related to an array of covariate- and statin-adjusted blood and urine biomarkers[24] (**Figure 3b**). These include phosphate (PC142 and PC133 — 6.3% and 3.9% — these components also have contribution from cholesterol medications) as well as urea and gamma glutamyltransferase (PC135 - 4.2%). Another relevant component (PC147 - 3.8%) has contribution from phosphate and hair color. All components

have contribution from variation proximal to the lipoprotein genes *LPA* and *APOC1*, along with variants at the *9p21.3* susceptibility locus (*CDKN2B*) and in the brain-expressed solute carrier *SLC22A3*[25] (**Supplementary Data 3**).

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[26]. The top three components (**Figure 3c**) for gout share strong contribution from covariate-adjusted blood urate[24]. One component (PC94 - 7.3%) is further driven by intraocular pressure and fat percentage; another (PC108 - 7.7%) is related to visual acuity (logMAR test results); and the third (PC117 - 5.9%) 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[27]. The transporter protein *ABCG2* is also key to both PC94 and PC108, and has been shown to play a role in renal urate transport[28]. 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[29].

Painting DeGAs Risk Profiles:

To further characterize the architecture of genetic risk for these traits, we "painted" the profiles of each high-genetic-risk individual (top 5% of dPRS), plotting a breakdown of each person's genetic risk across DeGAs components[15] which we call the DeGAs risk profile (**Methods**). This measure captures relevant underlying genetic diversity among high risk individuals (**Figure 4a-c**) in a way which complements the population-level scores from DeGAs. For example, the trait squared cosine score for BMI suggests that PC1, PC3, and PC169 are the top 3 components; but the DeGAs risk profile further implicates PC6 (with strong contribution from autoimmune traits) and PC7 (with strong contribution from standing height and body impedance) in genetic risk at the extreme tail of dPRS for BMI (**Figure 4a**).

We therefore investigated the diversity of driving components among high-risk individuals using their DeGAs risk profile. We used the Mahalanobis criterion (**Methods**) to find individuals in the entire test population whose risk profiles significantly differed from average. We then intersected these outliers (z-scored Mahalanobis distance > 1) with the high-risk individuals (top 5% of dPRS) to identify "high-risk outliers". This group (**Figure 4d-f**) has similar contributing components as the high-risk individuals (**Figure 4a-c**), but their relative importance to each of the individuals is quite different. This suggests that the DeGAs risk profile, as a personalized measure, can identify individuals with high genetic risk who are poorly described by "typical" trait pathology.

To better describe genetic diversity among these atypical individuals, we attempted to identify genetic subtypes of each trait in the high-risk outlier population. We performed a k-means clustering of this group using DeGAs risk profiles as the input; k was chosen using an iterative approach based on the marginal increase in variance explained resulting from incrementing the

number of clusters (**Methods**). We described each cluster using its mean risk profile (**Figure 4a-c**) and noticed that cluster membership divides individuals based on cPRS for relevant components (**Figure 4d-f**).

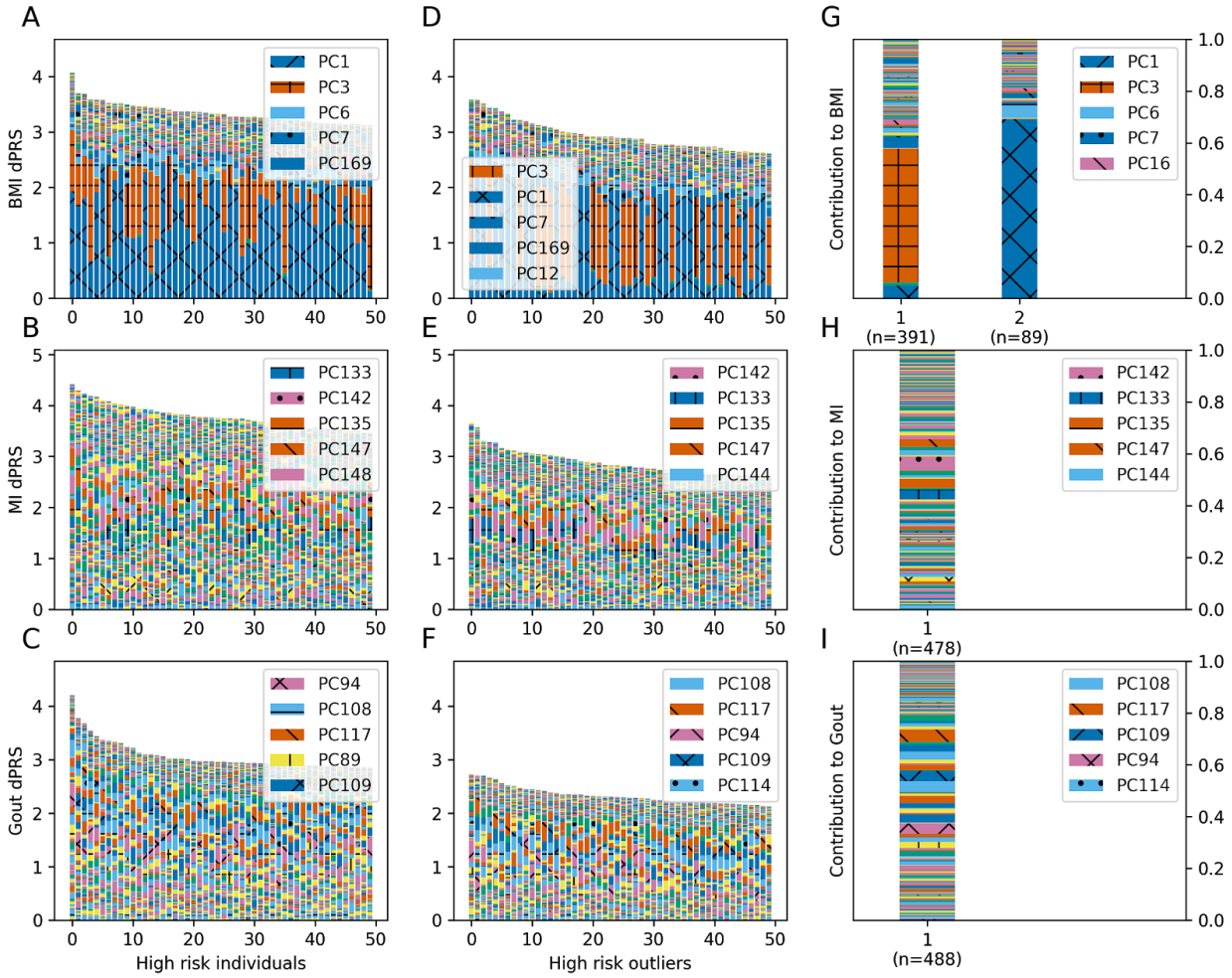

Figure 4: Painting components of genetic risk. (A-C) Component-painted risk for the 50 individuals or **(D-F)** outliers with highest dPRS for each trait in the test set. Each bar represents one individual; the height of the bar is the covariate-adjusted dPRS, and the colored components of the plot are the individual's DeGAs risk profile, scaled to fit bar height. Colors for the five most represented components in each box are shown in its legend in rank order. **(G-I)** Mean DeGAs risk profiles from *k*-means clustering of high risk outlier risk profiles, annotated with cluster size (*n*). Phenotype groups for selected components in this figure include: PC1 (Fat free mass); PC3 (Fat mass); PC142 (Phosphate and Hair color); PC133 (Phosphate and HCMV antigens); PC117 (Cystatin C and Urate); and PC108 (Urate and Visual acuity).

For body mass index, we identify two risk clusters (**Figure 4g**): one driven by the fat mass component (PC3 - 59.0%, *n*=264) and the other by the fat-free mass component (PC1 - 70.4%, *n*=20). 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 fat-free 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 risk for high BMI suggest alternative preventative and therapeutic approaches.

We find one cluster of risk for myocardial infarction, driven by components which all have strong contribution from blood phosphate and cholesterol lowering medications (**Figure 4h**). One component (PC142 - 5.4%) 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 - 4.0% and 3.4%) have contribution from blood (gamma glutamyltransferase, aurea) and urine (enzymatic creatinine, sodium) biomarkers[24], 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*[30].

There is also only one cluster of outliers for gout (**Figure 4i**), and the mean risk profile closely mirrors the driving components as ordered by the trait squared cosine score (PC94, PC108, and PC117). Of note is the increased importance of PC109, which is driven by platelet (thrombocyte) volume and width, covariate-adjusted blood urate[24], and brain MRI measures (T2-star) which can capture hemorrhaging phenotypes. Increased serum uric acid has been associated with cerebral microbleeds in stroke patients[31], which suggests this component captures shared biology related to 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.

Discussion:

In this study, we describe a novel approach (dPRS) to model polygenic traits using components of genetic associations. We build an example model using data from unrelated white British individuals in the UK Biobank to show that our method adds an interpretable dimension to traditional polygenic risk models by expressing disease, lifestyle, and biomarker-level elements in trait-related genetic components. Predicting genetic risk with these components led us to infer disease pathology beyond variant-trait associations with no loss of predictive power from reducing model rank (**Supplementary Figure 2**).

For three phenotypes of interest (BMI, MI, and gout) we showed that the DeGAs risk profile offers meaningful insight into the genetic drivers of trait risk for an individual. We then used this measure to identify clusters of high risk individuals with similar genetic load for each of the traits. We find, as in previous work[15], that genetic risk for BMI can be decomposed into fat-mass and fat-free mass related components. We also show that while many individuals have risk for BMI driven by a combination of the two components, there exist “outlier” individuals who have strong

contributions from only one of them. Our results further indicate that this diversity of contributory genetic risk is not limited to BMI, but extracting biological insights for other traits will likely require deeper phenotyping or other rich resources like single cell data.

We further demonstrated the generalizability of dPRS by assessing its performance in independent test sets of white British and non-British white individuals (**Supplementary Figure 4**; all traits in **Supplementary Table 1**) from the UK Biobank. Although we show dPRS is predictive in independent samples with some variability in ethnic composition, concerns about the generalizability of traditional clump-and-threshold PRS across groups also apply to our method. Though methods exist to identify suspected causal variants via fine-mapping, we decided to LD-prune variants prior to analysis with DeGAs. This approach is agnostic to patterns of association observed for particular phenotypes. However, this may leave dPRS slightly more vulnerable to overfitting patterns of LD in the GWAS population compared to other approaches, and may be worth revisiting in future work.

We also note that our analysis of subtypes may not be robust to different choices of input traits or study population. Taking gout as an example, our study finds only one cluster of outliers (**Figure 4c**) whose mean DeGAs risk profile mirrors its trait squared cosine score (**Figure 3i**). This absence of clusters may be due to urate acting by similar mechanisms to influence risk for gout across our cohort, but this may not be the case in other groups. Here, we excluded traits which may have noisy or confounded patterns of genetic associations: specifically, rare conditions ($n < 100$ in the UK Biobank) or traits which correlate with social measures like socioeconomic status. We encourage replication efforts using similar methods, and have made all DeGAs risk models from this work available on the Global Biobank Engine[32] (**Resources**).

We anticipate many potential applications of component-aware polygenic risk models like dPRS. Heritable conditions with known or putative biomarkers would be good candidates for follow-up studies that investigate an outcome jointly with its related quantitative features. For example, brain and liver images, metabolomics, and serum and urine biomarkers have been collected in resources like the UK Biobank, and may be of interest for future work[33,34]. Since DeGAs requires only summary-level data, it is possible to build a component model of genetic risk in one cohort (or across several) and use it to profile genetic risk and identify trait subtypes in another. Such analyses will help elucidate the diversity of polygenic risk for complex traits across individuals and populations.

Methods:

Study populations

The UK Biobank is a large longitudinal cohort study consisting of 502,560 individuals aged 37-73 at recruitment during 2006-2010[16]. 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.

We used a subsample consisting of 337,151 unrelated individuals of 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 report final associations and performance measures in the test population. An additional cohort of unrelated non-British Whites[24] (*n*=25,486) is used as an additional independent evaluation set. The “white British” and “non-British white” populations were defined using a combination of genotype PCs from UK Biobank’s PCA calculation[16] and self-reported ancestry (UK Biobank Field 21000) as a reference[24].

Genome-wide association studies in the UK Biobank:

PLINK v2.00a[35] [2 April 2019] was used for genome-wide associations of 805,426 directly genotyped variants, 362 HLA allelotypes, and 1,815 non-rare (AF > 0.01%) copy number variants[18] (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 using the --pheno-quantile-normalize flag. The following covariates were used: age, sex, the first four genetic principal components, and, for variants present on both of the 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 the UK Biobank[16]. 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 (thresholds varied by ancestry) or missingness (> 5%) were excluded from the data release. Prior to use in downstream methods, we performed additional variant quality control on array-genotyped variants, including more stringent filters on missingness (> 1%), gross departures ($p < 10^{-7}$) from Hardy-Weinberg Equilibrium, and other indicators of unreliable genotyping[36]. As with previous versions of by DeGAs, we further filtered variants by minor allele frequency (MAF > 0.01%) and LD-independence[15]. The LD independent set was computed with --indep 50 5 2 in PLINK[37] v1.90b4.4 [21 May 2017]. In total, this resulted in a set of 454,565 variants (452,639 genotyped variants, 130 HLA allelotypes, and 1,796 CNVs) for analysis.

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[36,38]. Additional traits like biomarkers, environmental variables, and self-reported questionnaire data like health outcomes and lifestyle measures, were collected from fields curated by the UK Biobank and processed using in-house methods[15,24]. 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. In all, we collected 1,905 traits with at least 100 observations (quantitative measures) or cases (binary traits); a full list of traits and their Global Biobank Engine phenotype IDs is in **Supplementary Table 1**. Summary statistics from all GWAS described here are publicly available on the Global Biobank Engine[32] (**Resources**). 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 summary statistics from GWAS computed using the above methods, we performed matrix Decomposition of Genetic Associations (DeGAs), as previously described[15]. First, a sparse matrix of genetic associations W ($n \times m$) was assembled using effect size estimates (or z-statistics) between n=1,905 traits and m=454,565 non-rare variants (MAF > 0.01%). Only variants with at least 2 nominally significant associations were used ($p < 0.01$; **Supplementary Figure 1** has additional cutoffs). After filtration, input statistics were standardized to zero mean and unit variance within each trait so as to weight them equally relative to one another.

We then performed a truncated singular value decomposition (TSVD) on W using the TruncatedSVD function in the scikit-learn python module[39,40] to identify the top c=500 trait-related genetic components. This factorization results in three matrices whose product approximates W: a trait singular matrix U ($n \times c$), a variant singular matrix V ($m \times c$), and a diagonal matrix S ($c \times c$) of singular values, which we denote by s_i for the i-th component (Figure 1a).

The matrices U, S, and V were then used to compute component polygenic risk scores (cPRS). The component PRS for the i-th DeGAs component can be written as

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

for an individual with genotype vector G ($m \times 1$) over the variants used in DeGAs. Here, $S_{i,*}$ denotes the i-th row of S. Using the cPRS for each component, we define the DeGAs polygenic risk score (dPRS) for the j-th trait as

$$dPRS_j = \boldsymbol{\Sigma}_i U_{j,i} \, cPRS_i$$

where $U_{j,i}$ is the (j,i)'th entry of U. In terms of the matrices U, S, and V, this can be expressed as

$$dPRS_j = U_{j,*} S V^T G$$

For interpretability, the population distribution of dPRS for each trait *j* is scaled to zero mean and unit variance, independent of the distributions of dPRS for other traits.

We further relate individuals to traits via components using a measure we call the DeGAs risk profile (dRP). An individual's profile for a given phenotype *j* is a vector over the *c* DeGAs components, where the value for the *i*-th component is proportional to

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

with a denominator introduced for normalization so that these values sum to one. To estimate the contribution of each component to an individual's overall genetic risk, we only consider component scores which have the same sign as the overall risk score (hence the max operator). This gives normalized risk profiles consisting of driving components for high-risk individuals with positive dPRS and protective components for low risk individuals with negative dPRS.

Computing polygenic risk scores:

As a baseline model for dPRS, we computed single-trait polygenic risk scores (PRS) with a pruning and thresholding approach using the same summary statistics used as input to DeGAs. These variants were already filtered on LD independence and filtered according to $p < p^*$ for some critical value p^* (see above), so no further processing was required. For a given DeGAs instance, the weights for prune- and threshold PRS for trait *j* were taken from the *j*-th row of *W*. The PRS was then 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; these are also the same assumptions used in 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 computed dPRS and DeGAs risk profiles for each trait using the above formulas.

Optionally, PRS and dPRS were adjusted by the following covariates: age, sex, and four genetic principal components from UK Biobank's PCA calculation. Adjustment was performed by fitting a multiple regression model with (d)PRS and covariates in the 10% validation population. We also fit a covariate-only model using the same procedure (without either polygenic score), and use its performance as baseline for the joint models (**Figure 2**).

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 filtering *p*-values for both betas and *z*-statistics. Performance of a DeGAs instance was assessed using the average correlation between its resulting set of dPRS models and their respective traits. For all traits used in the decomposition, we computed Spearman's rho (rank correlation) between dPRS and covariate-adjusted trait residuals in the validation population. Residual traits are the result of regressing out the following covariates: age, sex, and four genetic principal components. We find optimal performance in the validation population using *z*-statistics, a *p*-value cutoff of 0.01, and 300 components (**Supplementary Figure 1**).

For this final DeGAs instance, we present several assessment metrics for each polygenic score — dPRS and PRS with DeGAs data — within each study population (**Figure 2**). For each score and population, we estimated disease prevalence and mean quantitative trait values at various population risk strata. We also estimated the effect of (d)PRS quantiles on traits using a two step approach. First, in the training set we compute:

$$Y \sim \beta_0 + \beta_1 sex + \beta_2 age + \sum_{i=1}^{4} \beta_{i+2} PC_i$$

Then, in the test/validation set we estimate the effect β due to PRS quantile using the above parameters like so:

$$Y \sim \hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 age + \sum_{i=1}^{4} \hat{\beta}_{i+2} PC_i + \beta 1_{PRS(q)}$$

where $\mathbf{1}_{PRS(q)}$ is an indicator function which is 1 if an individual is in the quantile of interest *q* (e.g. 0-2%) and 0 if the individual is in the baseline group (40-60% quantile of (d)PRS). Individuals in neither the quantile of interest nor the baseline group were excluded; if individuals were in both *q* and the baseline group (e.g. if *q* were 45-40%) they were counted in *q* and removed from baseline.

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* 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 covariates. As baseline, we also report AUC for a covariate-only model (see above).

Classifying genetic risk profiles from DeGAs components:

In order to assess whether our method could identify subtypes of genetic risk, we analyzed the DeGAs risk profiles of high-risk individuals whose dPRS is driven by an "atypical" combination of DeGAs components. We used the Mahalanobis criterion (D_M) to identify outlier individuals whose *z*-scored distance from the mean DeGAs risk profile exceeded 1:

$$D_M = \sqrt{(x-\mu)S^{-1}(x-\mu)^T}$$

where *x* is the DeGAs risk profile; μ is the mean profile; and *S* is the identity matrix. Traditionally *S* is taken to be the covariance matrix for each of the features across all *x*'s: we model each of the components as having equal variance so as to identify "atypical" individuals rather than statistical outliers. We note that this formula reduces to the Euclidean distance between a DeGAs risk profile *x* and the mean profile μ.

We then intersected this set with the top 5% of dPRS values to create the "high risk outlier" group. Here, we define the mean risk profile for a trait as the component-wise mean across all individuals' DeGAs risk profiles in high risk set (top 5% of dPRS).

To identify subtypes among high risk outliers, we performed a *k*-means clustering of their DeGAs risk profiles using the KMeans function from the python scikit-learn module[39]. The number of clusters *k* was determined iteratively using proportional reduction in error. We iteratively incremented *k* and recomputed 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 each cluster, across all clusters — failed to exceed 25% on the *k*+1st clustering. We then evaluated which components drove risk in each cluster by computing a mean risk profile for the group (defined as above), and renormalized it to one for visualization (**Figure 4g-i**).

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 in the 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. M.A.R. is supported by Stanford University.

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., and M.A.R. carried out quality control of the data. R.T. and T.H. aided in statistical design and conception. We thank Johanne Justesen, Michael Wainberg, and members of the Rivas Lab for comments on the manuscript. 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.

Resources:

Supplementary data, including weights for the final DeGAs model, are available on the Global Biobank Engine[32]: https://biobankengine.stanford.edu/downloads.

References:

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

17. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

18. 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).

19. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

20. 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).

21. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40**, 575–583 (2008).

22. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).

23. 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).

24. Sinnott-Armstrong, N. *et al.* Genetics of 38 blood and urine biomarkers in the UK Biobank. doi:10.1101/660506

25. 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).

26. 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).

27. 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).

28. 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).

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

30. 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).

31. 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).

32. McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty999

33. Yu, B. *et al.* The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies. *Am. J. Epidemiol.* **188**, 991–1012 (2019).

34. Fest, J. *et al.* Search for Early Pancreatic Cancer Blood Biomarkers in Five European Prospective Population Biobanks Using Metabolomics. *Endocrinology* **160**, 1731–1742 (2019).

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

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

37. 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).

38. DeBoever, C. *et al.* Assessing digital phenotyping to enhance genetic studies of human diseases. doi:10.1101/738856

39. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

40. 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:

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

Figure S2: Summary of final DeGAs instance — validation versus test R; scree plot

Figure S3: Individual-level concordance of dPRS and PRS for BMI, MI, and gout.

Figure S4: dPRS performance for BMI, MI, and gout in UK Biobank non-British white individuals.

Table S1: Counts of individuals at intersecting quantiles of risk by PRS and dPRS for BMI, MI, and gout.

Data S1: List of phenotypes and (d)PRS performance metrics across population groupings. This includes N, adjusted beta for the 2% strata of risk (BETA2), AUC of covariate-adjusted (d)PRS (binary only), Pearson correlation between (d)PRS and quantitative trait values (PEARSON_R), and Spearman rank-correlation between (d)PRS and trait value (SPEARMAN_R). Details on the computation of these measures and descriptions for each of the analysis populations are described in the main text (**Methods**).

Data S2: Phenotype contribution scores for all 300 DeGAs components.

Data S3: Gene contribution scores for all 300 DeGAs components. For these plots, we use the gene contribution score as previously defined[15]. Briefly, the contribution score a gene is considered to be the sum of contribution scores for each variant present in the gene. Noncoding variant are treated as singular “gene” entities.

Note: Supplementary Data are available on the Global Biobank Engine (**Resources**).

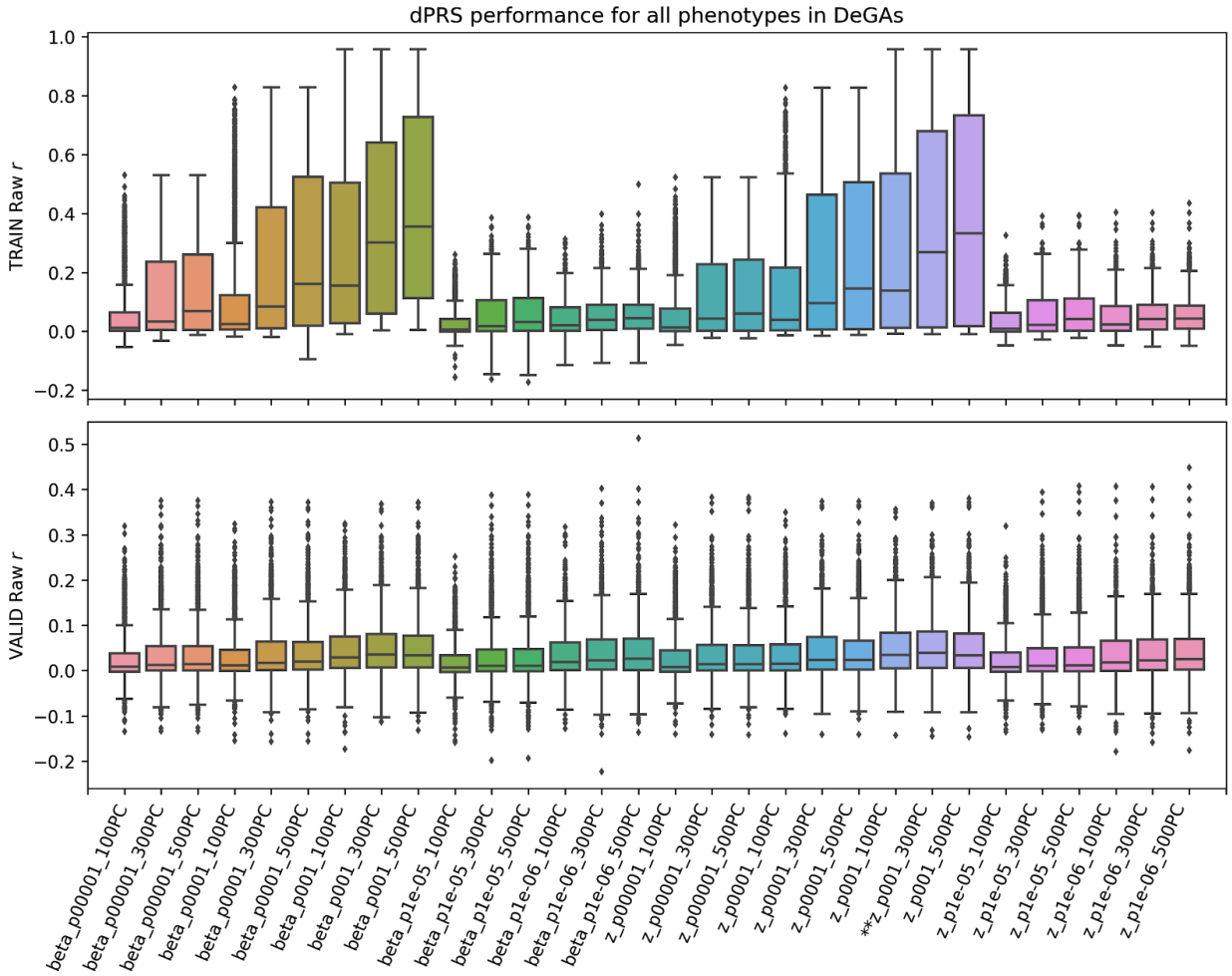

Figure S1: Hyperparameter optimization. Distribution of Spearman's rho between trait dPRS and trait values (top, in the training set; bottom, in the validation set) for all traits in several DeGAs models. We computed DeGAs across an array of parameters, varying input statistics (betas or z-statistics) from GWAS; minimum p-value filters (p=1e-2, 1e-3, 1e-4, 1e-5, 1e-6); and the number of components to compute (c=100,300,500). The model with optimal performance was chosen by maximizing mean rho between dPRS and trait in the validation set (bottom). It used z-statistics from associations with p<1e-2 and 300 components, and is labeled with two stars (**).

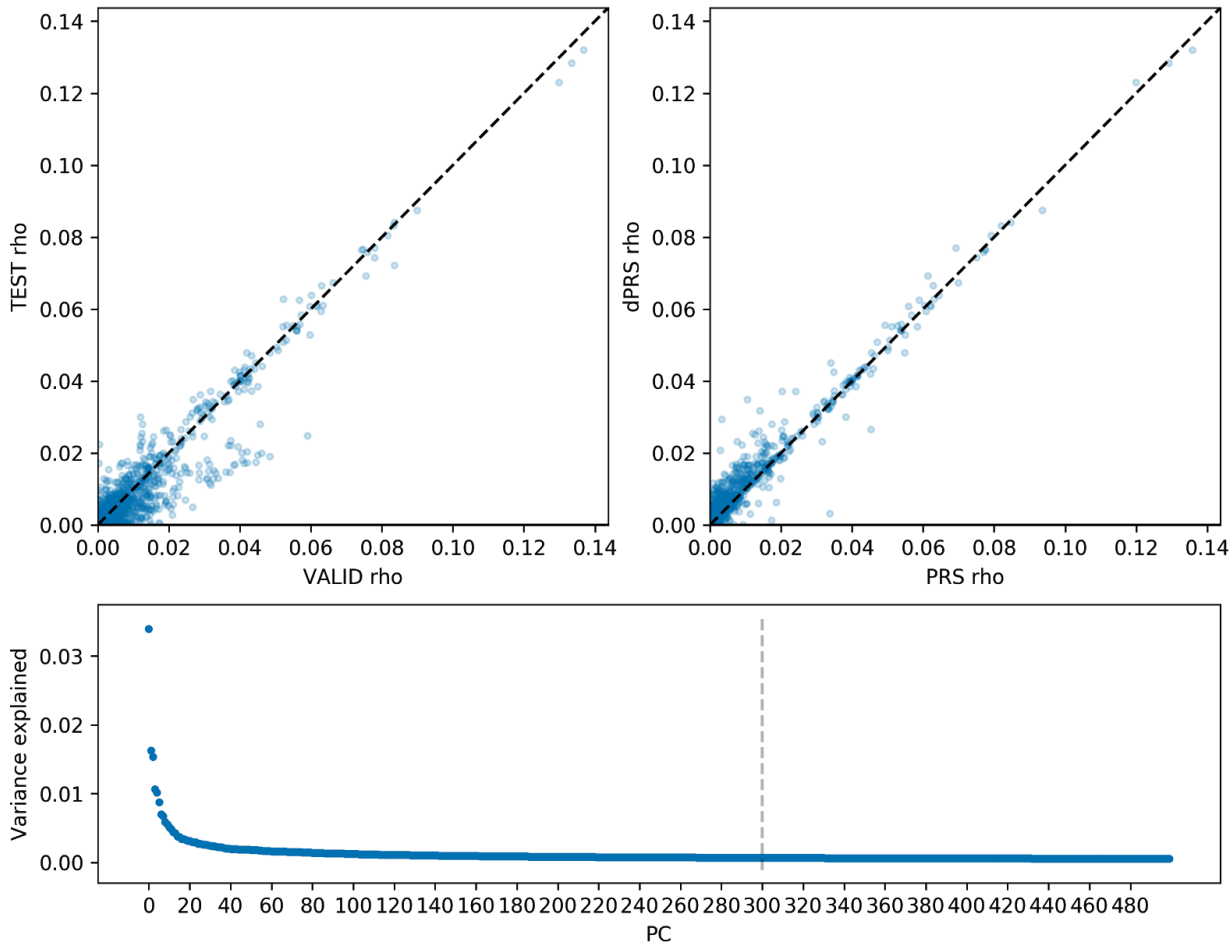

Figure S2: Performance of the final DeGAs model with optimized hyperparameters. We plot squared rank correlation (Spearman's rho) between dPRS and trait values across all traits in the test (top left) and validation sets (top left), or between dPRS and PRS in the test set (top right). The fraction of variance explained by each DeGAs component is shown in a scree plot (bottom). The grey dashed line is at the final *c*=300 components, which collectively explain ~47.4% of variance in the original input matrix.

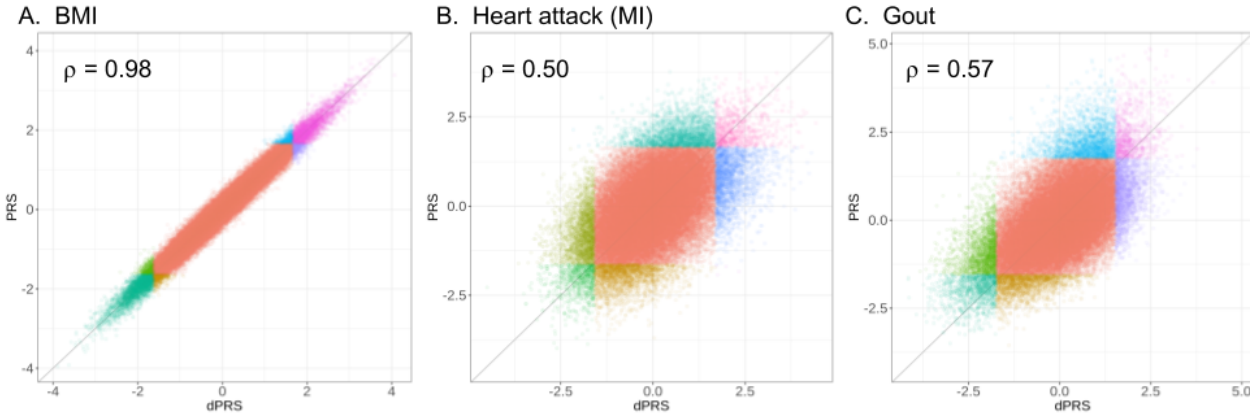

Figure S3: Correlation between dPRS and PRS. For BMI (A), MI (B), and gout (C), dPRS (x-axis) and PRS (y-axis) for all individuals in the test set are shown alongside rank correlation (Spearman's ρ) between the two. The gray diagonal line is y = x. Color represents whether the individuals are identified as high-risk (top 5%) or low risk (bottom 5%) by dPRS, PRS, or both (Supplementary Table X).

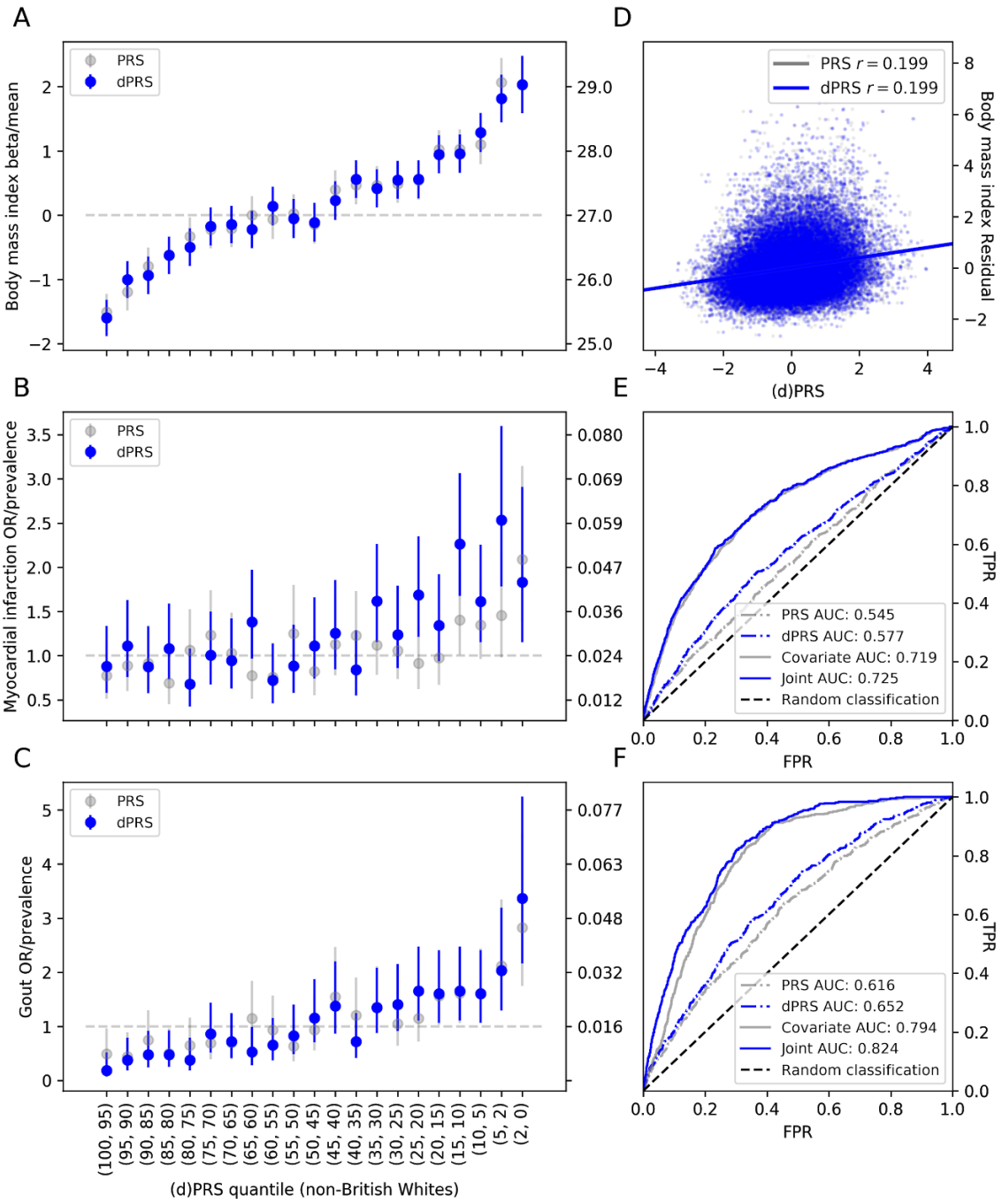

Figure S4: (d)PRS performance in non-British whites. Analog of **Figure 2** in a population of non-British white individuals (*n*=24,908) from UK Biobank. The second percentile of risk for dPRS (PRS) has: 2.06 kg/m^2 (2.06) higher BMI, 1.83-fold (2.09) increased odds for MI, and 3.37-fold (2.83) increased odds of gout, adjusted for age, sex, and 4 genetic PCs. Overall model performance of (d)PRS adjusted for these covariantes is measured by Pearson's *r* for BMI (D) or AUC for dPRS versus PRS alone, or dPRS + covariates versus a covariate model for the binary traits (E,F).

Supplementary Table 1. Comparison of risk stratification by dPRS and PRS. For the three traits we highlight in our study, we count the number of individuals in the same (or different) risk strata under each model. Binary cases and controls are further split ("phenotype" column) within each bin of risk for dPRS and PRS. Counts are shown for the top 5% (A) and bottom 5% (B).

A

dPRS	PRS	BMI	MI	Gout	Phenotype
Top 5%	Top 5%	2830	46	48	Case
			829	840	Control
	Bottom 95%	531	134	105	Case
			2362	2378	Control
Bottom 95%	Top 5%	531	120	83	Case
			2376	2400	Control
	Bottom 95%	63346	2148	1084	Case
			59414	60491	Control

B

dPRS	PRS	BMI	MI	Gout	Phenotype
Bottom 5%	Bottom 5%	2799	16	4	Case
			803	1224	Control
	Top 95%	563	66	5	Case
			2487	2139	Control
Top 95%	Bottom 5%	563	68	30	Case
			2485	2114	Control
	Top 95%	63313	2298	1281	Case
			59206	60632	Control