PLOS GENETICS

RESEARCH ARTICLE

Reverse GWAS: Using genetics to identify and model phenotypic subtypes

Andy Dahl[1]*, Na Cai[2,3], Arthur Ko[4], Markku Laakso[5,6], Päivi Pajukanta[4], Jonathan Flint[7], Noah Zaitlen[1]*

1 Department of Medicine, UCSF, San Francisco, California, United States of America, **2** Wellcome Sanger Institute, Cambridge, United Kingdom, **3** European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom, **4** Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California, United States of America, **5** Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, Finland, **6** Kuopio University Hospital, Kuopio, Finland, **7** Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, UCLA, Los Angeles, California, United States of America

* andywdahl@gmail.com (AD); noah.zaitlen@ucsf.edu (NZ)

Citation: Dahl A, Cai N, Ko A, Laakso M, Pajukanta P, Flint J, et al. (2019) Reverse GWAS: Using genetics to identify and model phenotypic subtypes. PLoS Genet 15(4): e1008009. https://doi.org/10.1371/journal.pgen.1008009

Editor: Gregory S. Barsh, Stanford University School of Medicine, UNITED STATES

Received: October 5, 2018

Accepted: February 7, 2019

Published: April 5, 2019

Data Availability Statement: Genotype data from the CONVERGE cohort are available through the European Variation Archive (EVA) as previously described [https://www.ncbi.nlm.nih.gov/pubmed/28195579], and phenotype data are available through the CONVERGE data access committee (http://www.well.ox.ac.uk/converge). Genotype and phenotype data from the METISM cohort are available through METSIM data access committee (http://www.nationalbiobanks.fi/index.php/studies2/10-metsim).

Abstract

Recent and classical work has revealed biologically and medically significant subtypes in complex diseases and traits. However, relevant subtypes are often unknown, unmeasured, or actively debated, making automated statistical approaches to subtype definition valuable. We propose reverse GWAS (RGWAS) to identify and validate subtypes using genetics and multiple traits: while GWAS seeks the genetic basis of a given trait, RGWAS seeks to define trait subtypes with distinct genetic bases. Unlike existing approaches relying on off-the-shelf clustering methods, RGWAS uses a novel decomposition, MFMR, to model covariates, binary traits, and population structure. We use extensive simulations to show that modelling these features can be crucial for power and calibration. We validate RGWAS in practice by recovering a recently discovered stress subtype in major depression. We then show the utility of RGWAS by identifying three novel subtypes of metabolic traits. We biologically validate these metabolic subtypes with SNP-level tests and a novel polygenic test: the former recover known metabolic GxE SNPs; the latter suggests subtypes may explain substantial missing heritability. Crucially, statins, which are widely prescribed and theorized to increase diabetes risk, have opposing effects on blood glucose across metabolic subtypes, suggesting the subtypes have potential translational value.

Author summary

Complex diseases depend on interactions between many known and unknown genetic and environmental factors. However, most studies aggregate these strata and test for associations on average across samples, though biological factors and medical interventions can have dramatically different effects on different people. Further, more-sophisticated models are often infeasible because relevant sources of heterogeneity are not generally known *a priori*. We introduce Reverse GWAS to simultaneously split samples into homogeneous subtypes and to learn differences in genetic or treatment effects between

Funding: This work was funded by 377 National Institutes of Health (NIH) grants 1U01HG009080-01, 5K25HL121295-03, 378 1R03DE025665-01A1, HL-095056, HL-28481, and U01 DK105561. N. Cai was supported 379 by the EBI-Sanger Postdoctoral Fellowship. A. Ko was supported by the NIH grant 380 F31HL127921. 381 The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

subtypes. Unlike existing approaches to computational subtype identification from high-dimensional trait data, RGWAS accounts for covariates, binary disease traits and, especially, population structure–important features of real genetic datasets. We validate RGWAS by recovering known genetic subtypes of major depression. We demonstrate RGWAS can uncover useful novel subtypes in a metabolic dataset, finding three novel subtypes with both SNP- and polygenic-level heterogeneity. Importantly, we show that RGWAS can uncover subtypes with differential treatment response: we show that statin, a common drug and potential type 2 diabetes risk factor, may have opposing subtype-specific effects on blood glucose.

Introduction

Distinguishing subtypes can be essential for treatment, prognosis, and learning basic disease biology. For example, breast cancer has subtypes distinguished by tumor hormone receptor status that have different genetic risk variants, population structure, comorbidities, treatment responses, and prognoses [1, 2]. Many other common diseases have known, biologically distinct subtypes [3–7], often involving distinct tissues or biological pathways, including two diseases we study: major depression (MD) [8] and type 2 diabetes (T2D) [9, 10]. Genetically distinct subtypes can arise from gene-environment interactions [11–13]; gene-gene interactions [14, 15]; or disease misclassification, which is well documented but usually ignored [16, 17].

In this work we describe a novel method to learn and validate genetic subtypes in a two-step approach that we call reverse GWAS (RGWAS). In the first step, RGWAS infers subtypes by clustering multiple traits with a finite mixture of regressions method we designed specifically for large, multi-trait GWAS datasets (MFMR). The core assumption of MFMR is that the subtypes differ in distribution for many traits, which creates a subtype structure that can be detected with computational algorithms. In the second step, RGWAS assesses the causal *biological* distinction between the inferred subtypes by testing for genetic effect heterogeneity across subtypes, both at the SNP- and polygenic-levels. We also test effect heterogeneity for non-genetic covariates, like medical interventions, to assess *pragmatic* distinctions between subtypes [18–21].

RGWAS offers several important advances in both the identification and testing steps over other recent computational approaches to uncover subtypes [22–31]. First, unlike previous methods, the RGWAS identification step corrects for population structure, handles binary traits, and scales to tens of thousands of samples. Second, previous approaches to the testing step were badly confounded and/or under-powered to detect genetic heterogeneity in complex traits, while RGWAS offers calibrated *p*-values, polygenic subtype validation tests, and covariate adjustment. These advances substantially reduce both type I and type II errors in computational subtyping.

We first evaluate RGWAS through extensive simulations over a wide range of parameter settings and generative models, including several scenarios that violate our assumed model. In comparison to previous methods, as well as several novel methods and obvious extensions, we find that RGWAS is substantially more powerful, more robust, and better calibrated. We then validate RGWAS in real data by recovering a recently discovered stress subtype in MD, as well as known subtype-specific SNP effects.

Finally, we apply RGWAS to a metabolic cohort where subtypes are unknown *a priori* and find strong statistical support for genetic effect heterogeneity across dozens of complex

phenotypes: first, a subtype-aware mixed model substantially increases heritability, from 20.7% to 30.2% on average across traits; second, we identify dozens of subtype-specific SNP effects that could not be discovered in standard analyses, including three SNPs with previously identified metabolic interactions; third, subtype-aware GWAS increases the number of hits, from 60 to 70 across traits. Crucially, we find that statin, a widely prescribed drug that may increase diabetes risk [32–34], has opposing effects on blood glucose across metabolic subtypes, which suggests that learned subtypes may have significant translational value.

Results

Reverse GWAS is calibrated and powerful in simulations

We examine the relative behaviors of RGWAS and other recent methods for computational subtype identification through application to simulated datasets. We simulated from a range of generative models and parameter settings meant to reflect many of the complexities of real data (Methods, Supplementary Section 3). Our baseline includes noise that is correlated across traits; large main subtype effects; 27 quantitative and 3 binary traits; and 12 SNPs containing null, homogeneous, and heterogeneous SNPs. We simulate datasets both under a $K = 2$ model and a $K = 1$ model where no subtypes of any sort are present. For each parameter setting we aggregate results from roughly 300 simulated datasets. We explore several variations to this baseline later.

For each simulated dataset we first cluster individuals (RGWAS step 1) into a subtype vector z. We focus on our Multi-trait Finite Mixture of Regressions (MFMR) approach (Methods) and three other, conceptually distinct approaches to cluster individuals into subtypes (Methods). First, we use Gaussian Mixture Models (GMM) to represent covariate-unaware methods, e.g. k-means [28, 31] and TDA [22, 26, 27]. Second, we consider a novel approach based on Canonical Correlation Analysis (CCA) that defines the subtype vector z as the top phenotypic CC. Third, we use the true z to show the best-case scenario with perfect subtyping (Oracle).

First, we observed that MFMR outperformed non-oracle methods in recovering true cluster identities across a range of simulations settings (S1 Fig). While this does increase confidence in MFMR, clustering accuracy is not a directly useful metric because true clusters are not known in practice. Moreover, we are primarily concerned with the significance and interpretation of our inferred subtypes, rather than merely their existence, which requires calibrated tests for effect heterogeneity between identified clusters. Hence our primary evaluation focuses on the false- and true-positive rates (FPR and TPR) for the standard SNP-subtype (z) interaction test in (3), conditioning on main subtype effects (RGWAS step 2). A polygenic alternative for RGWAS step 2 is presented below (Methods, [35]). We report a true positive if the SNP was simulated with heterogeneous effects across subtypes, and a false positive if the SNP was simulated with null or homogeneous effects.

The results in Fig 1 show that interaction tests applied to MFMR subtypes are calibrated and almost perfectly obtain oracle power. Crucially, MFMR remains calibrated even when $K = 1$ (Fig 1b), so RGWAS discoveries reliably validate the existence of subtypes. Further, when $K > 2$ subtypes were simulated, MFMR with fixed $K = 2$ lost power but remained calibrated (S2a Fig). Together, this shows that MFMR is robust to misspecified K, though power increases when K is more accurate.

On the other hand, GMM is miscalibrated by an order of magnitude when $K = 1$, making it unreliable for interaction testing and, more broadly, validating the existence of subtypes (Fig 1b). One fundamental difference between MFMR and GMM is that only MFMR models covariates. Intuitively, ignored covariate effects confound true subtypes, hence GMM performs poorly when homogeneous effects are large or heterogeneous effects are small (Fig 1c, S2c–S2e Fig).

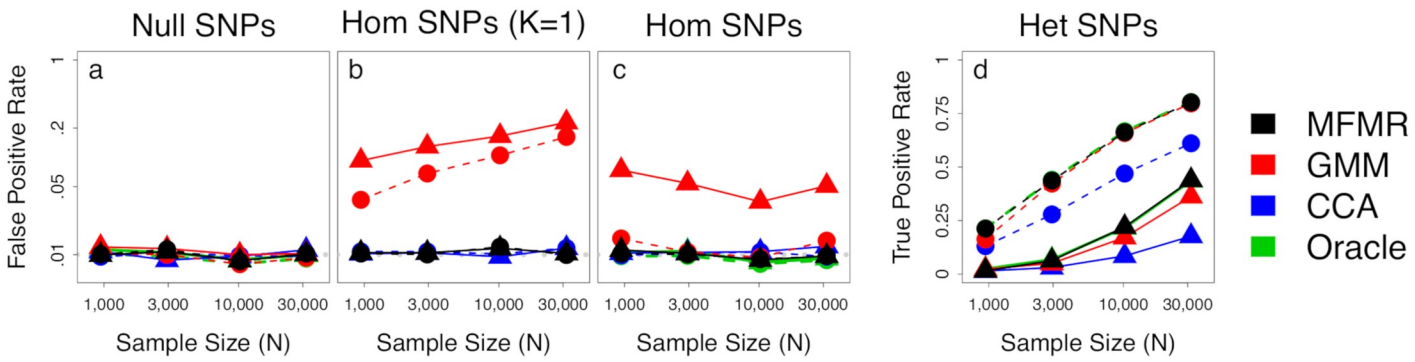

Fig 1. SNP heterogeneity tests at nominal $p = .01$. SNPs are either null (a); homogeneous (Hom, c); or truly heterogeneous (Het, d); we also test Hom SNPs in simulations with no subtypes (b). FPR is shown on the log scale. Hom SNPs explain 4% of variance and Het SNPs explain .4% (triangles) or vice versa (circles).

https://doi.org/10.1371/journal.pgen.1008009.g001

Another difference between MFMR and GMM is that the former explicitly models binary traits, while the latter discards them and, concomitantly, loses power (S2b Fig, Methods).

Our novel application of CCA has low power but seems calibrated and, sometimes, seems to outperform even the oracle (e.g. S2c and S3 Figs). Unsurprisingly, this is a Pyrrhic victory: by smoothing over traits, CCA creates a bias that simultaneously increases FPR and TPR, which we show with theory (Supplementary Section 4) and simulation (S3 Fig). Moreover, the CCA subtypes poorly capture the true, discrete subtypes (S1 Fig). Broadly, CCA is often calibrated for testing the existence of heterogeneity, but it cannot determine which specific traits are truly heterogeneous and, moreover, has lower power than MFMR. We also tried using the top phenotypic PC for z, which performed like CCA except with lower power, and the top genetic PC [25], which had very low power (S4 Fig).

Simulations violating model assumptions. Having analyzed simulations from our assumed model, we turn to other generative models and parameter settings to assess the limitations of RGWAS. We first considered several direct variations to our baseline simulation. We drew t_5-distributed noise, which has higher kurtosis than the Gaussian noise assumed by our model. We found that MFMR and CCA performed similarly as in the Gaussian case, but GMM became even more inflated (S5 Fig); nonetheless, we caution that MFMR, like linear regression, will not generally be calibrated under arbitrary noise distributions. Next, we assessed non-linear SNP effects and found that MFMR again remained roughly calibrated, but GMM and CCA were inflated for strong homogeneous effects; this held for two types of non-linear covariate transformations (S6 Fig), but we again caution that MFMR is not generally calibrated when covariates have non-linear effects. We also tried replacing the discrete subtypes z with a continuous z drawn from a Gaussian distribution. This exacerbated GMM inflation, but MFMR remained powerful and calibrated; however, when we doubled all heterogeneous effect sizes, MFMR became $\approx$ 4-fold inflated for large homogeneous covariates and $N \geq 30,000$ (S7 Fig), suggesting caution for MFMR in large datasets when large-effect, continuous subtypes may exist. Overall, modest amounts of these types of model misspecification cause relatively modest RGWAS bias, but stronger model violations will inevitably cause FPR inflation.

Second, we examined case ascertainment, which causes bias in many genetic testing settings [36–40]. We simulated "Case/Control" studies of a relatively rare disease, with population prevalence 20%, by ascertaining a 50/50 case/control dataset (S8 Fig). Little changed qualitatively, though all methods, even the oracle, had slight FPR inflation for large N. This modest level of inflation is expected because the ascertainment process violates the interaction regression model [41].

Third, we examined the robustness of MFMR to covariate/phenotype misclassification. Running MFMR requires specifying whether each available variable is a trait or a covariate (though SNPs are always covariates because of Mendelian randomization, and all the covariates in our simulation are SNPs). Intuitively, subtypes are clusters of covariate-adjusted traits. Variables that confound subtype structure, then, should included as covariates, while variables that may have different distributions between subtypes should be included as traits. Nonetheless, this distinction can be murky in practice, and so we perform simulations where we treat a covariate like a trait or vice versa. MFMR remained calibrated when swapping traits and covariates, unlike GMM, suggesting exact covariate/trait specification is not essential for valid inference with MFMR. However, MFMR did lose power, emphasizing the statistical utility derived from properly modelling the distinction between covariates and traits (S9 Fig).

Fourth, we studied the impact of SNP-subtype correlation (Supplementary Section 3.2), which is a known source of bias in genetic interaction tests [42]. We found that power held roughly constant for all methods as the G-E correlation increased from 0 to 1, but FPR did increase for all methods, including the oracle (S2f Fig). Nonetheless, the MFMR inflation is modest–always less than 3-fold–and MFMR was no more inflated than the oracle. Altogether, this simulation does recapitulate known biases from G-E correlation, but it does not suggest that RGWAS is more susceptible to these biases than standard tests for genetic interaction.

Finally, and most importantly, we examined the robustness of the clustering methods to population structure, which is routinely the primary confounder in genetic studies. We simulated a 50/50 mixture of two populations and 10,000 SNPs from a Balding-Nichols model with $F_{ST} = .1$ (Supplementary Section 3.3). We repeated our simulations using 12 randomly selected SNPs (out of the 10,000) and adding population main effects of varying strength. For MFMR and the oracle, we condition on three genetic PCs and their interactions with z in the step 2 tests. MFMR remains calibrated and powerful while CCA and GMM suffer substantial FPR inflation, even for completely null SNPs (S8 Fig). Using PCs in step 2 largely addresses the inflation from using GMM in step 1, but only when subtype structure is stronger than population structure; otherwise, the inferred GMM subtypes are non-trivially confounded by population structure. This is important because population structure is often stronger than subtype structure in reality, especially in multi-ethnic datasets, and because GMM produces false positives when subtypes are completely absent. Altogether, successful genetic subtype inference requires population structure adjustment both when inferring subtypes in step 1 and when testing their genetic significance in step 2.

Positive control: A known major depression subtype

The results over the simulated datasets showed that RGWAS was powerful and calibrated across a wide range of parameter settings. To see if RGWAS could perform well in a real setting, we used CONVERGE [43], a major depression (MD) cohort with a recently discovered, genetically heterogeneous "Stress" subtype [44]. This serves as a positive control for genetic subtype discovery. In addition to having a known subtype, CONVERGE is ideally suited to RGWAS analysis because it recruited a large number ($N = 9,303$) of deeply phenotyped (31 binary and 10 quantitative traits) Han Chinese women with recurrent MD and matched controls. Cases were carefully ascertained to minimize environmental heterogeneity, comparatively amplifying signals for biological heterogeneity. RGWAS analysis is also compelling because only a small number of genetic associations with MD have been found to date–consistent with the existence of genetically distinct disease subtypes–and there are few known genetic interactions in complex human traits.

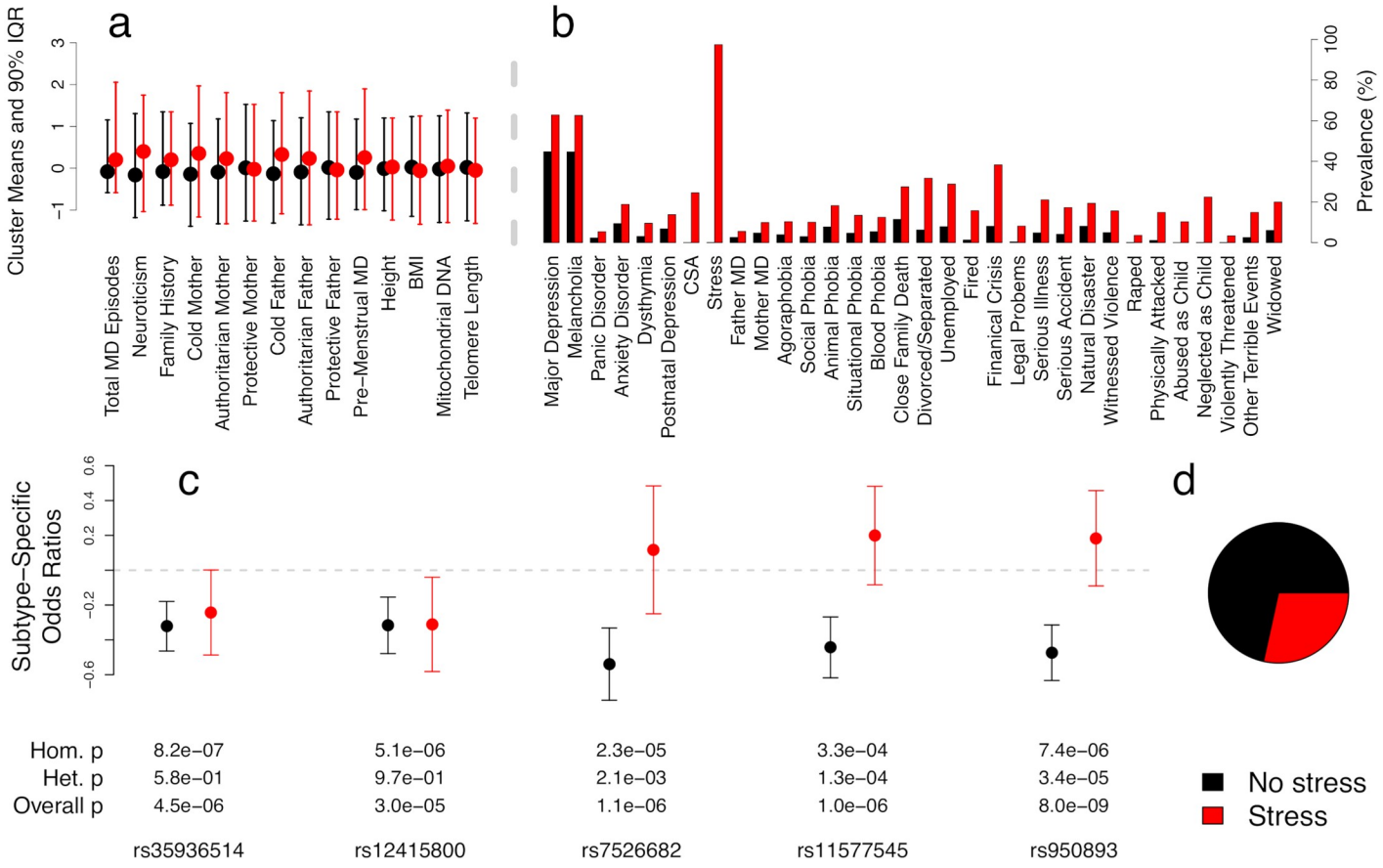

Fig 2. Genetic heterogeneity in the CONVERGE major depression dataset. Quantitative trait 90% inter-quantile ranges (a) and binary trait prevalences (b) are shown for each subtype. (c) Per-subtype odds ratios (±2 s.e.) for two SNPs discovered by (homogeneous) GWAS [43] (left) and three SNPs discovered using known subtypes [44] (right).

https://doi.org/10.1371/journal.pgen.1008009.g002

The inferred MFMR subtypes (RGWAS step 1) with $K = 2$ are summarized in Fig 2. We conditioned on age and ten genetic PCs as covariates, and we jointly imputed the covariates and traits (Methods). The MFMR method split the individuals into subtypes that distinguish the aggregate lifetime adversity measure "Stress", which recovers the subtype chosen by domain experts [44]. While "Stress" is an obvious contributor to MD risk, there is no reason to expect MFMR would identify this as the key factor given we are studying a large number explicitly MD-relevant traits. Indeed, GMM did not split along this trait (squared correlation with "Stress" is .01), demonstrating "Stress" is not a trivially obvious subtype. The GMM clusters were not examined further as they have high FPR in simulations.

We tested five SNPs for effect heterogeneity (RGWAS step 2) across subtypes (Fig 2c) with our standard SNP-subtype interaction test. The first two SNPs (rs35936514 and rs12415800) were discovered in the initial GWAS [43] and we use as negative controls for heterogeneity. For positive controls, we use three SNPs (rs7526682, rs11577545, and rs950893) that we recently found to interact with "Stress" [44]. As expected, the homogeneous SNPs are nearly genome-wide significant, and RGWAS successfully determines which of the five SNPs are heterogeneous. Note that the heterogeneous SNPs show only modest homogeneous signal because they essentially have no effect in the "Stress" subtype. In previous work, we established that these subtypes have differential polygenic architecture using linear mixed models

($p = .038$); we also found suggestive polygenic score interaction and suggestive differences in the respective heritability estimates per subtype [44], further supporting the genetic distinction between "Stress" subtypes of MD.

We chose $K = 2$ using prior knowledge that MD can be split by (binary) "Stress". We assessed this empirically by evaluating the MFMR likelihood on held-out data, which supported $K > 1$ subtypes (S10 Fig). $K = 3$ creates an MD-only subtype, so we do not pursue $K \geq 3$.

New metabolic subtypes with genetic and pragmatic significance

We next applied MFMR (RGWAS step 1) to metabolic traits measured in METSIM [45]. By combining genetic, environmental, metabolomic, and disease measurements, METSIM enables tracing the pathway from risk factors to metabolic consequences to altered disease risk. We studied 6,248 unrelated Finnish men. We used three binary traits: 854 samples had T2D, 3,526 had pre-diabetes (preT2D), and 541 had coronary heart disease (CHD); we excluded 15 samples with T1D. We used 13 broad quantitative traits and 228 nuclear magnetic resonance (NMR) metabolite measurements. We projected the NMR traits onto their top 6 PCs (capturing 77% of variance), which is a standard way to ease computation while retaining most of the information in the raw NMR traits. As covariates, we used three genetic PCs, age, age^2, and smoking, alcohol, statin, diuretic, and beta-blocker use.

We study $K = 3$ to compromise between parsimony and the cross-validated log-likelihood, which broadly supports $1 < K \leq 7$ (S10 Fig). To test robustness to perturbations of the data, we used five-fold cross-validation. We found that 99.6% of originally co-clustered pairs (i.e. same likeliest subtype) remained together, showing that people from the same training population can be accurately assigned to existing subtypes.

The three inferred metabolic subtypes are summarized in Fig 3. They primarily distinguish the metabolomic PCs, which are aggregates of 228 NMR traits. To elucidate the subtypes, we fit logistic regressions on the raw NMR traits, conditional on statin, and studied those with nominal $p < .01$ (despite [28, 30], these p-values are not calibrated). We first compared the

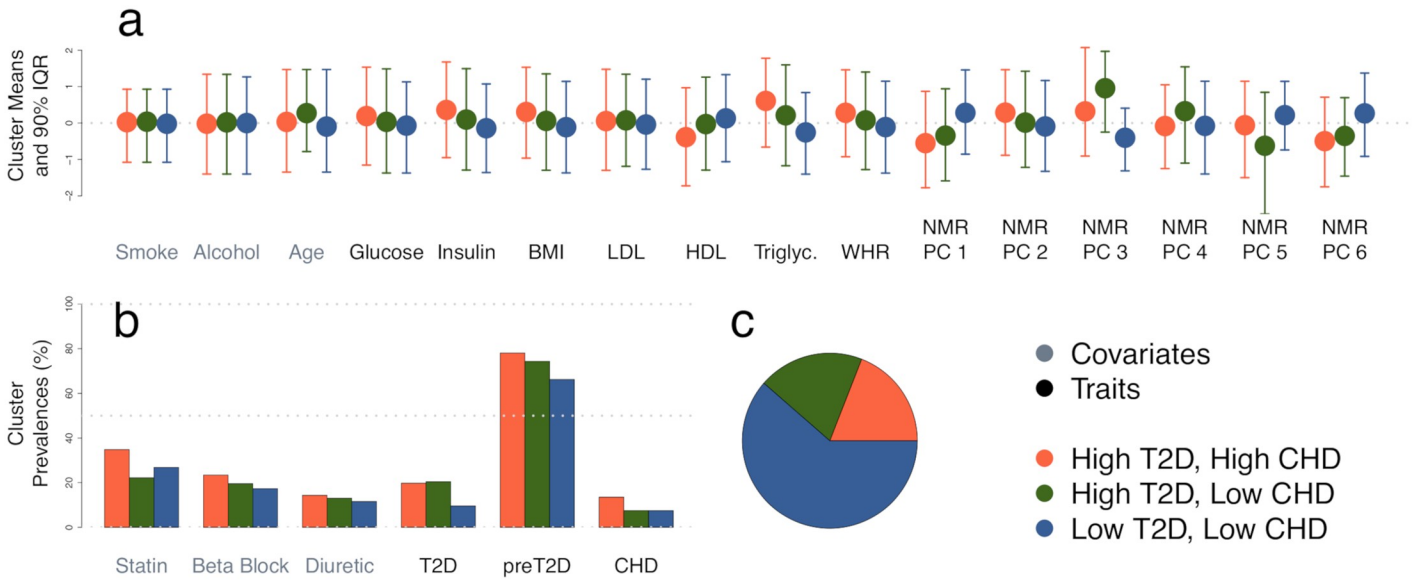

Fig 3. Three inferred metabolic subtypes in METSIM. (a) Quantitative and (b) binary distributions for covariates (grey labels, left) and traits (black labels, right). (c) Subtype sizes.

https://doi.org/10.1371/journal.pgen.1008009.g003

large blue group to the combined orange and green groups, which suggested the blue group had less-esterified cholesterol in small HDL and higher histidine and relative amounts of omega-3 fatty acid. Next, comparing orange to green indicated orange had more free but less esterified overall cholesterol, especially in large LDL, and that orange has higher polyunsaturated fats and phenylalanine.

SNP level metabolic heterogeneity. We next sought to evaluate the subtypes for evidence of SNP effect heterogeneity on metabolic phenotypes (RGWAS step 2). However, with ∼6,000 samples, we do not have power for genome-wide heterogeneity tests. Instead, we test known metabolic GWAS SNPs–68 from T2D and 13 from CHD (Methods). We found four heterogeneous SNP-trait associations at $p = .05/81$ (Fig 4). The orange and green effect estimates had opposite sign for 2/4, and all blue estimates were near zero. This suggests that the blue group is a type of baseline and that partially overlapping biological pathways are specifically activated in the smaller groups.

These SNPs have several known metabolic interactions, providing additional evidence that the subtypes are meaningful. rs10401969 is a splice variant for *SUGP1* that affects downstream splicing in the gene targeted by statins, *HMGCR* [46]; it also interacts with an APOE SNP on fenofibrate response [47]. rs7138803 interacts with exercise for obesity [48] and features in an obesity score interacting with diet [49]. rs780094 interacts with another SNP for fasting glucose [50], suggestively interacts with diet [51], and is one of three SNPs in a risk score interacting with postprandial and post-fenofibrate cholesterol [52].

We additionally performed heterogeneity tests of these 81 SNPs directly over the 228 raw NMR phenotypes we used to construct the NMR PCs. We found 33 further SNP-trait interactions at $p = .05/81$ (S11 Fig). This included interactions between rs7138803 and 8 VLDL traits and insulin; rs7528419 and 16 VLDL traits and HDL; and rs780094 and triglyceride proportion in large VLDL. These results add confidence and interpretability to the interactions we discovered for NMR PCs 3 and 6. Further, we discovered 5 additional SNPs, not significantly heterogeneously associated with our 16 primary traits, including rs10885122 (interacting with medium HDL phospholipid percentage and HDL-3), rs1387153 (phospholid percentage in large and extra large VLDL), and rs6903956 (HDL).

To show an example of the utility of identifying subtypes, we performed a genome-wide interaction scan with the global, *K* df test [53]. This "GxE GWAS" test does not establish SNP heterogeneity, but it can increase power over ordinary GWAS when heterogeneity exists. GxE GWAS and GWAS give largely consistent results (Table 1, S12 Fig), as expected because the homogeneous and global tests are not independent. Nonetheless, GxE GWAS is a valuable complement to GWAS as it discovers 10 additional loci (though GxE GWAS misses 19/60 GWAS loci).

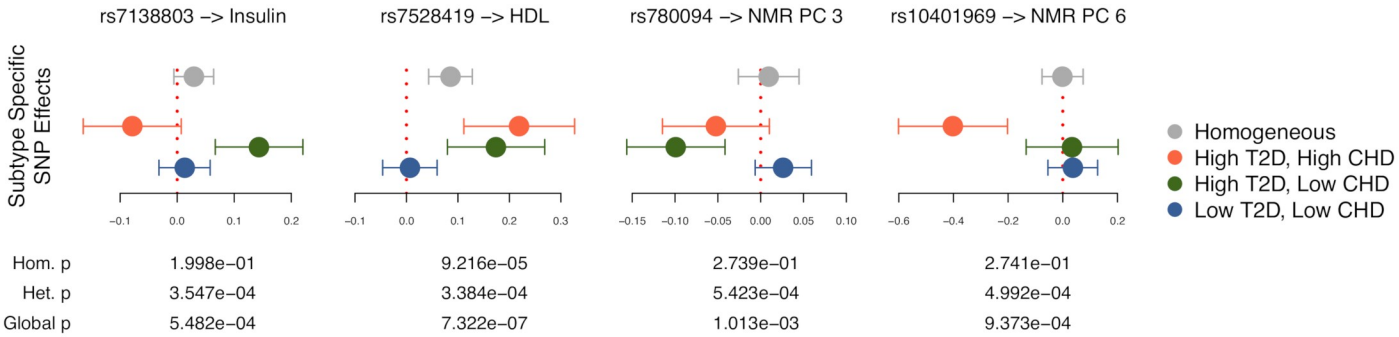

Fig 4. Metabolic subtype-specific SNP effects across the 16 traits used for subtyping. Subtype-specific effect estimates are shown ± 2 s.e. for significantly heterogeneous SNPs out of 81 known metabolic SNPs ($p = .05/81$).

https://doi.org/10.1371/journal.pgen.1008009.g004

Table 1. Number of genome-wide significant loci for GWAS and GxE GWAS. Shared describes the number of loci that are significant in both the GWAS and GxE GWAS for the same trait ($r^2 < .2$), while GWAS and GxE GWAS describe the loci unique to that association scan. No loci were found for CHD, insulin, or WHR. Both approaches found a single preT2D locus. We excluded NMR PC 5 because of GxE GWAS inflation.

	Gluc	BMI	LDL	HDL	TG	PC 1	PC 2	PC 3	PC 4	PC 6
GWAS	1	0	3	2	6	5	2	0	3	4
GxE GWAS	2	1	2	1	2	1	3	1	1	1
Shared	2	0	5	5	3	1	2	0	3	19

https://doi.org/10.1371/journal.pgen.1008009.t001

To mimic prior approaches, we performed a covariate-unaware GxE GWAS by using GMM subtypes (in step 1) and by excluding covariates from the heterogeneity tests (in step 2). Genome-wide tests were highly inflated for $K \in \{2, 3, 4\}$, usually obtaining effectively infinite λ_{GC} (S1 Table). Notably, λ_{GC} was even inflated for the binary traits, which were excluded from the clustering in step 1 in order for GMM to converge, emphasizing the subtlety and breadth of overfitting concerns in two-step testing. This inflation can easily be mistaken for strong, ubiquitous signal when evaluating only candidate SNPs, which is common in computational subtyping papers. The RGWAS λ_{GC} (using MFMR in step 1 and covariate-aware tests in step 2) were comparatively modest, with a maximum of 1.33 (after excluding NMR PC 5, with $\lambda_{GC} = 1.83$). Despite this modest inflation for some traits–which can be readily detected, and avoided, by evaluating λ_{GC}–RGWAS results are usable, unlike the covariate-unaware results that mimic existing approaches. Similar conclusions hold for the global, K df test.

Polygenic metabolic heterogeneity. Identification of heterogeneous effects at individual SNPs provides both evidence of differential causal effects between subtypes and the specific loci that distinguish them. To complement these results, we additionally employed a polygenic linear mixed model (LMM) test for genetic subtype heterogeneity (RGWAS step 2). Such polygenic approaches have greater power to detect genetic signal than SNP-level tests, but they do not identify individual causal loci.

Briefly, the polygenic model estimates both the ordinary heritability and the subtype-specific heritability, the latter aggregating all subtype-specific SNP effects. For example, when subtypes are two sexes, the subtype-specific heritability aggregates all SNP effects that are active only in one sex (and, more generally, SNP effects that differ between sexes). For non-sex linked traits, the resulting sex-specific heritability is zero, hence nonzero subtype-specific heritability demonstrates the existence of subtype-specific genetic effects.

We fit the subtype specific LMM with IID GxEMM [35], which is a simple extension of a previous LMM for GxE [54] that accommodates probabilistic subtype membership. This is important because hard-assignment of individuals to subtypes discards information and fails to propagate step 1 subtyping uncertainty to step 2 subtype testing. We use the covariates used in our MFMR decomposition as fixed effects, as well as their interactions with subtype status [55]. We calculate variance explained after residualizing fixed effects [56]. We test for $h^2_g > 0$ with a Wald test, and we test for $h^2_{het} = 0$ with an LRT. For T2D, we exclude people with preT2D; for preT2D, we exclude people with T2D.

We report estimates for binary traits on the observed scale as there is no commonly accepted approach for rescaling heterogeneous heritability estimates to a liability scale. Similarly, we do not report p-values for these binary traits, to be maximally conservative. Nonetheless, we do observe that IID GxEMM increases the estimated heritability for preT2D and CHD, which is consistent with the existence of subtype-specific heritability.

We found significant subtype-specific heritability for 4 of the 13 quantitative traits used in the MFMR decomposition ($p < .01/13$, Fig 5, left), giving strong evidence that our inferred subtypes tag meaningfully distinct biology. On average, these subtype-aware heritability

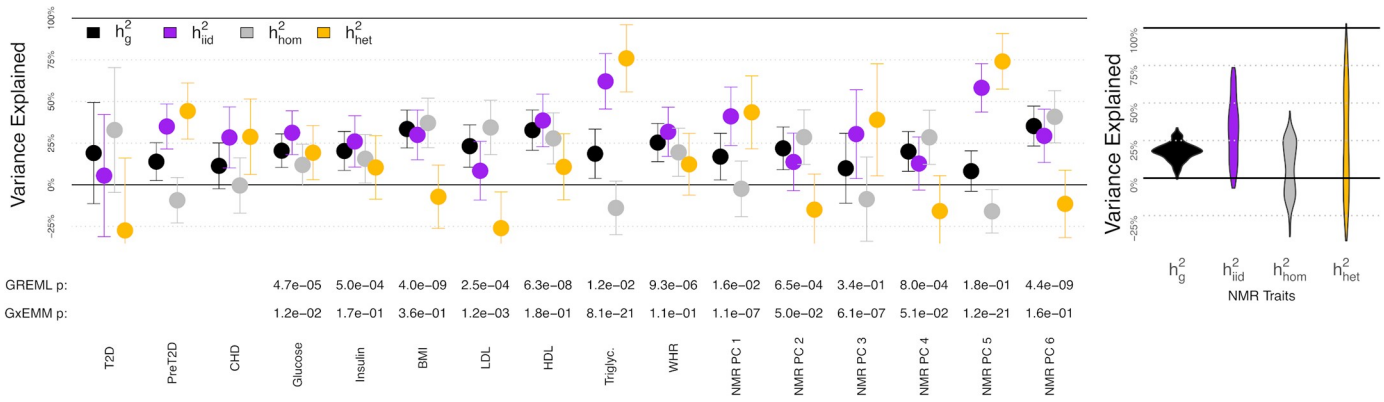

Fig 5. Polygenic heterogeneity in the inferred metabolic subtypes. Left: Point estimates ± 2 s.e. for traits used in MFMR clustering. Right: Across-trait estimate distribution for all 228 raw NMR traits. h^2_g is the standard heritability estimate from GREML [54]. h^2_{hom} and h^2_{het} are the homogeneoues and subtype-specific heritability estimates from IID GxEMM, and h^2_{iid} is the sum of h^2_{hom} and h^2_{het}.

https://doi.org/10.1371/journal.pgen.1008009.g005

estimates are 30.2%, compared to 20.7% for ordinary heritability estimates (S13 Fig). This shows that subtypes can mask substantial heritability across an array of traits. Intuitively, this increase in heritability derives from allowing subtype-specific effects that are ignored by homogeneous heritability: in the sex example, a SNP with exactly opposite effects in males and females has zero net contribution to homogeneous heritability, but it clearly contributes to the broader notion of subtype-specific heritability.

To expand to a larger set of traits and bolster confidence in widespread subtype-specific heritability, we repeated our IID GxEMM analysis on the 228 NMR traits (Fig 5, right). We found significant subtype-specific heritability for 104 of the 228 NMR traits ($p < .01/228$), dramatically increasing the number of significantly polygenically heterogeneous traits. On average, the subtype-aware heritability is 31.8% and the standard heritability is 17.5% (Fig 5), which are qualitatively comparable to the average heritabilities from the 13 quantitative MFMR traits.

Pragmatic metabolic heterogeneity. While identifying genetic heterogeneity is important for showing that subtypes have differential causal biology, identifying nongenetic sources of subtype heterogeneity can be pragmatically important. We tested for statin effect heterogeneity to assess the ability of the metabolic subtypes to differentiate medical intervention effects. Using our standard interaction test with $K = 3$ (Methods), only glucose had significant statin heterogeneity at $p = 0.05/16$ ($p = 1.19 \times 10^{-4}$). There is no obvious FPR inflation as statin has no other significantly heterogeneous effects across other traits (S2 Table). These results held after conditioning on T2D as a heterogeneous covariate in our step 2 test (statin-glucose interaction $p = 2.95 \times 10^{-6}$). Statin effect heterogeneity is further supported by tests with $K = 4$ (T2D-adjusted heterogeneity $p = 9.67 \times 10^{-7}$); for $K = 2$, however, the test is insignificant ($p = .39$), providing further evidence that $K = 2$ gives insufficient subtype resolution.

We next estimated the subtype-specific statin effects with our primary metabolic subtypes (derived treating statin as heterogeneous inside MFMR). We performed a heterogeneous linear regression on glucose conditional on our standard covariates and T2D, which indicated that statin increases blood sugar in most people–consistent with [32, 33]–but also that it may decrease glucose in the smaller, higher-risk orange and green groups.

Since METSIM measured two time points, we next tested the pragmatic ability of our baseline subtypes to predict conversion from preT2D to T2D. We fit logistic regression on time 2 T2D status for the 1,924 baseline prediabetics with time 2 data. Subtypes significantly

predicted T2D conversion ($p = 0.0036$), with orange and green converting less often than blue, which remained after conditioning on our standard covariates ($p = .031$). This also demonstrates the subtypes persist at least partially over time, in contrast to prior, directly age-dependent T2D subtypes [57].

Discussion

In a purely *descriptive* sense, inferring subtypes is easy: applying any clustering algorithm to any data produces subgroups. But existing methods cannot go beyond such descriptions because they are liable to downstream FPR inflation. By contrast, RGWAS is calibrated in simulation, recovers known MD subtypes, and produces *biologically* and *pragmatically* validated metabolic subtypes. RGWAS handles covariates, mixed binary and quantitative traits, and residual trait correlations.

There are several limitations to RGWAS. First, like other two step methods, RGWAS fails to propagate first-step uncertainty. Similarly, although we do not imagine there is a "true" K, more can always be done to better choose K. Also, while we have tested a variety of simple decompositions to learn subtypes, others may perform better, especially where domain-specific tools exist. In particular, MFMR is conceptually similar to a matrix factorization/depth-two linear network, suggesting inner layers of appropriate neural networks may define useful subtypes.

There are also specific limitations to our inferred stress subtypes in CONVERGE. First, our stress measurements were retrospective and self-reported, hence they may be biased by MD status. Second, our analysis was not entirely without domain supervision because we included the aggregate trait "Stress" that was previously manually constructed [44]. Nonetheless, RGWAS identified the key trait amongst dozens, unlike GMM, and our METSIM analysis demonstrates that RGWAS can be useful without any domain guidance.

The statin effect heterogeneity on glucose we found is consistent with previously reported interactions between statin and age [32] and genetically predicted LDL [34] on T2D, and also fenofibrate's interaction with lipid levels on cardiovascular risk [58]. By contrast, large meta-analyses did not find inter-study statin heterogeneity [32, 33]. This suggests that statin heterogeneity largely exists within, rather than between, datasets. In the future it will be important to more precisely characterize the causal mechanism underlying statin heterogeneity. Although we expect such a mechanism to broadly replicate, the same is not necessarily true for our specific subtypes, especially as they rely on phenotypic PCs which are dataset-specific.

MFMR is only a first step toward genetic subtyping, and there are many possible extensions. Sparsifying penalties can be incorporated by replacing CM steps with calls to third-party software and could extend MFMR to higher-dimensional traits and covariates. A random-effect version of MFMR could improve power to detect polygenic subtypes, though computational issues are non-trivial. MFMR could also be adapted to count data, zero-inflation, higher-order arrays, or missing data. Theoretically, it would be interesting to let subtypes vary between traits, which MFMR can capture only non-parsimoniously by choosing large K. Instead of an i.i.d. prior on z, MFMR could model z with a multinomial logistic regression to estimate, test, and correct for effects on z, which can be directly interesting [59]; this can also be important for correcting bias from G-E correlation [42]. Or, instead, we could use a continuous prior on z with a factor analysis model [23]. (We note, however, that RGWAS copes reasonably well with modest G-E correlation and continuous z in our simulations.) Finally, MFMR could be applied only within diseased individuals to directly define subtypes of disease, though this requires fundamentally different step 2 tests (c.f. [60]).

Our polygenic approach to subtype validation with GxEMM provides a much needed power advantage over SNP-level heterogeneity tests at the cost of resolution; conceptually,

polygenic risk score tests lie between [61, 62]. But SNP-level precision is not needed to meet our criterion for biologically meaningful subtypes, making GxEMM invaluable for subtype validation. Nonetheless, there are many limitations to our approach that we will address in future work. First, GxEMM can confuse non-linear effects for heterogeneity. Similar issues arise in generalized linear models, where the existence of effect heterogeneity depends on the choice of link function. Some forms of non-linearity can be accommodated [63], including ascertained case/control data where cases are over-sampled from the population [36–40]; we are working to extend these approaches to leverage subtype heterogeneity. Second, IID GxEMM assumes that each subtype has equal heritability and equal noise levels, which may not hold in practice for many meaningful subtypes. Several approaches exist to relax these assumptions to varying degrees [35, 64, 65], which reduce bias, provide richer characterizations of subtypes, and will be important to pursue in future work. However, this increased generality comes with higher estimation error for individual parameters, and we consider the results presented in our paper conservative alternatives to these richer models.

Although we focused on fixed- and random-effect interaction tests to establish heterogeneity between subtypes in RGWAS step 2, it may also be useful to apply recent, complementary heterogeneity tests. For example, Subtest could be used to assess differences between $K = 2$ disease-only subtypes [60]. For large K, on the other hand, StructLMM is a natural complement to GxEMM: the latter is more powerful because it uses genome-wide information and a richer GxE model, but the former has SNP-level resolution and scales to dramatically larger N and K. Similarly, large-K subtypes could be post-processed with hierarchical clustering and tested with TreeWAS [62]. Broadly, any heterogeneity test can be used in the second step.

In the future, we will actively encourage MFMR to prioritize tissue-specific subtypes by incorporating tissue-specific genetic risk scores as heterogeneous covariates. This is particularly interesting for traits, including metabolic diseases [15, 66], that have disparate genetic risk factors acting through distinct cell types, tissues, or biological processes. Subtypes that differentiate biological modes of action at this systems-level would be more easily interpretable and useful for basic research and precision treatment. In larger datasets, it may also be interesting to evaluate subtype-specific enrichments in heritability explained per tissue or cell type [67].

Finally, as MFMR seeks clusters that are unaffected by confounders like population structure, age, or sex, it may be useful for clustering in settings where protecting certain information is important for privacy or fairness [68]. In this sense, MFMR is to GMM roughly as AC-PCA [69] or contrastive PCA [70] are to ordinary PCA.

Methods

Ethics statement

CONVERGE: The study protocol was approved centrally by the Ethical Review Board of Oxford University (Oxford Tropical Research Ethics Committee) and the ethics committees of all participating hospitals in China. All participants provided written informed consent.

METSIM: The Ethics Committee of the University of Eastern Finland and Kuopio University Hospital approved the METSIM study, and this study was conducted in accordance with the Helsinki Declaration. All participants provided written informed consent.

RGWAS step 1: Clustering with MFMR to find subtypes

Reverse GWAS first clusters samples into subtypes (step 1) and then tests for covariate effect heterogeneity between subtypes (step 2). RGWAS always uses MFMR in step 1. Step 2 always

features interaction tests between subtype membership and focal covariates, which may be non-genetic, a SNP, or all SNPs in the genome.

We derive a novel clustering algorithm, multitrait finite mixture of regressions (MFMR), beginning from the standard regression model for interaction. Assuming a single quantitative trait y, covariates X, discrete subtypes z, and a focal covariate g putatively interacting with z, our model is:

$$y_i = X_i, \alpha + \gamma_{z_i} + g_i \beta_{z_i} + \epsilon_i \quad (1)$$

X_i, is a vector of Q control covariates, like genetic PCs or sex, with homogeneous effect sizes α. $z_i \in \{1, \ldots, K\}$ is a K-level factor specifying the subtype for individual i, and γ_k are the subtype main effects. β is the vector of subtype-specific g effect sizes. We say g is homogeneous if $\beta_1 = \ldots = \beta_K$; otherwise, g is heterogeneous. We assume ϵ is i.i.d. Gaussian with mean zero.

Our full MFMR model generalizes (1) in several complementary directions. First, we learn the subtypes (z) rather than assume they are known (giving a Finite Mixture of Regressions, FMR) by assuming z_i are i.i.d. Categorical:

$$P(z_i = k|p) = p_k \quad \text{for } k = 1, \ldots, K \quad (2)$$

Second, we generalize the single trait y to a matrix Y of Multiple traits (MFMR), which adds power for subtypes that affect the distribution of many traits. This power is important in practice because genetic interactions are often weak but phenotypic relationships are often strong. We also generalize the single heterogeneous covariate g to a matrix G of multiple covariates, which adds power when there are many heterogeneous effects.

Finally, we model binary traits with probit link functions to mitigate the spurious local modes that plague methods like k-means. For example, this issue led others to discard roughly half their data *post hoc* [31]. The full multi-trait probit model is computationally prohibitive even for modest B, which we address with a novel conditional independence assumption. This induces constraints in our optimization which we solve analytically with block matrix identities (Supplementary Section 2.3).

Computationally, we fit MFMR with an Expectation Conditional-Maximization (ECM) algorithm. Our ECM generalizes standard EM for Gaussian Mixture Models. Both iterate between z updates in E-steps and parameter updates (e.g. α and β) in (C)M steps.

When fitting MFMR in step 1, a covariate that will be tested for heterogeneity in step 2 can either be ignored (MFMRX), included in X (MFMR, our default), or included in G (MFMR+). In Gaussian mixture models, covariates can only be ignored (GMM) or added as traits (GMM+) [31]. MFMR+ and GMM+ overfit in simulations, inflating the FPR; conversely, MFMRX and GMM underfit homogeneous covariates, which also inflates FPR (S4 Fig). MFMR strikes a balance: the homogeneous effect is adjusted but subtypes are not tuned to the heterogeneous effect. This resembles a score test as the alternate is tested by evaluating only the null. Nonetheless, small-effect covariates, like SNPs, can be safely ignored within MFMR, enabling genome-wide testing with MFMRX [71]. For simplicity, we refer only to MFMR throughout our paper, but we use the MFMRX test for real human SNPs (in CONVERGE and METSIM, but not the simulations) as they likely have negligible effects; however, we use the MFMR test for statin in the main text because it has large metabolic effects.

We note that MFMR generalizes several well-known models. If binary traits and X are excluded and the covariates G are reduced to an intercept, MFMR becomes GMM. When $P = 1$ and z is known, MFMR becomes a standard gene-environment interaction (GxE) model with discrete environments/subtypes. Finally, if $P = 1$ and $\beta_k = \beta_0$ for all k, MFMR reduces to linear/probit regression.

RGWAS step 2: Calibrated tests to validate subtypes

For simplicity, we assume there are $K = 2$ subtypes and just one interacting covariate, g. These assumptions mean the output from step 1 is just a vector z, where z_i is the subtype 1 probability for sample i, and that the interaction model takes a simple form:

$$y \sim \tilde{X}\alpha + z\gamma + g\delta + (g * z)\beta + \epsilon \quad (3)$$

$\tilde{X}$ collects all background covariates, like genetic PCs, unlike existing subtype validation tests that largely ignore population structure [24, 26, 29, 31]. $*$ is element-wise multiplication, but it can easily be generalized to allow $K > 2$ and a matrix G instead of a single covariate g.

We consider three tests for g: the homogeneity test for $\delta \neq 0$ given $\beta = 0$; the heterogeneity test for $\beta \neq 0$ with free δ; and the global test for $\delta, \beta \neq 0$ [53]. The homogeneity test has 1 degree of freedom (df), the heterogeneity test has $K - 1$ df, and the global test has K df. We focus on the heterogeneity test, which establishes that g has differential effects across subtypes and thus that the subtypes differ in causal biology (if g is genetic) or pragmatically (e.g. if g is a treatment). We also investigate the global test's ability to increase power over the typical homogeneous test in the GxE GWAS analysis in Table 1. We assume ϵ is i.i.d and test with linear or logistic regression.

Polygenic step 2 tests. We also developed a polygenic version of the interaction test in (3) to jointly model and test δ and β across all SNPs with random effects. When there are no interaction effects, i.e. $\beta = 0$, this gives exactly the standard GREML approach to estimate heritability from genome-wide similarity across unrelated samples [56]. GREML accomplishes this by modelling each SNP's homogeneous effect, δ, as a small Gaussian variable, and then aggregating the size of each SNP's δ across the genome to estimate the total genetic contribution to phenotypic variability, i.e. heritability.

Polygenic interaction models go further by also giving the interaction effect, β, a random effect distribution. As the homogeneous mixed model used in GREML aggregates the δ estimates across the genome to estimate homogeneoues heritability, the interaction mixed model aggregates the β estimates along the genome to estimate the heterogeneous heritability explained by subtype-/environment-specific genetic effects. The latter provides an estimate of subtype-specific heritability, which can be combined with the homogeneous heritability estimates to partition broad-sense heritability into shared- and subtype-specific components. This approach was pioneered for unrelated humans in [54], but this model assumes that each sample is deterministically assigned to a single environment/subtype. Because our subtype assignments are probabilistic, we use GxEMM to fit subtype-specific heritabilities [35], which accommodates arbitrary environmental covariates.

Polygenic interaction tests are essential for genetic subtyping at modest sample sizes because the test for nonzero subtype-specific heritability is much more powerful than testing individual SNPs in complex traits. Polygenic interactions can demonstrate that subtypes have partially distinct genetic bases even when power is too low to discover individual subtype-specific SNP effects.

Other approaches to infer subtypes

We develop a novel subtyping approach by applying CCA to G and the joint binary and quantitative phenotype matrix (Y^b: Y), each column-wise centered and scaled, and taking z to be the top phenotypic CC. CCA (and phenotypic PCA) defines z as a linear trait combination. We prove that this causes the interaction tests to have inflated FPR when a mixture of heterogeneous and homogeneous traits are studied (Supplementary Section 4), which is likely in

practice. Nonetheless, sparse estimators can resolve this problem in some theoretical settings, and CCA is computationally efficient (S1 Fig).

We also tested GMM, which models samples as draws from one of K multivariate Gaussians. We fit GMM to the quantitative traits with a standard EM algorithm [72]; in early tests where binary traits were included, GMM often failed to converge, or converged to exactly coincide with one of the binary traits, even with multiple random restarts. We consider GMM similar, in the sense of covariate-unawareness, to k-means, which struggles even more with binary traits, and TDA, a proprietary package.

Most similar to MFMR, LIMMI aims to identify GxE with unknown E in gene expression [23]. Beyond many technical differences, LIMMI and MFMR are built for disjoint scenarios: MFMR only fits tens of traits, but LIMMI only fits hundreds of samples, preventing its use in our setting.

METSIM dataset

We selected metabolically relevant SNPs by taking published GWAS SNPs for T2D or CHD. We used the 153 T2D SNPs in Table 1 of [73] as known T2D SNPs. We had genotyped 86 of these SNPs, which we reduced further to 68 roughly independent SNPs ($r^2 < .1$). We used the 65 CHD SNPs in S2 Table of [74] as known CHD SNPs, 13 of which we genotyped (all $r^2 < .1$). We filtered the original 10,070 person dataset so all pairwise kinships were below 0.05, as in [56].

Phenotype imputation

We imputed missing data before running MFMR in CONVERGE. We jointly imputed covariates and traits with a sample-wise i.i.d. Gaussian model (MVN-impute from [75]). We thresholded imputed entries in Y^b to {0, 1} in order to retain the downstream logistic regression framework. By contrast, discarding samples with any missing data reduces sample size by roughly half and the known positive SNP interactions were no longer recovered.

We imputed METSIM similarly, including all 228 NMR traits at the imputation step. We used softImpute to accommodate the wide matrix [76].

We note that complete-data analyses performed in similar contexts substantially reduce sample size, e.g. [31] discard 39% of their samples.

Code availability

RGWAS is implemented in the simple, free `rgwas` R package, available with a vignette at https://github.com/andywdahl/rgwas.

All summary data and code necessary to reproduce the main and supplementary figures and tables are available at: https://github.com/andywdahl/rgwas-scripts.

Supporting information

S1 Text. Supplementary note. Full description of MFMR model and EM algorithm. Also includes full descriptions of simulations and proofs about PCA/CCA subtype estimators. (PDF)

S1 Fig. Running time and subtype estimation accuracy in simulations. Left: Average running times in main Fig 1 (excluding failed GMM runs). Right: Clustering accuracy for simulations without ('Quantitative', as in main Fig 1) and with ascertainment ('Case/Control', as in S2 Fig). We measure accuracy with adjusted Rand index, which varies from 0 (random guessing) to 1 (exact match). We compute the index only across pairs from a random 300

subsamples, reducing computation roughly $\approx 10^5$-fold when $N = 100,000$. Accuracies are estimated for roughly 300 simulations per point in the plot. MFMR+ is shown for simplicity because MFMR gives different clusters per tested SNP.
(TIF)

S2 Fig. Simulations varying several further parameters. Tests for truly heterogeneous SNPs are shown in the top 6 panels (a-f), and the corresponding tests for SNPs with only homogeneoues effects are shown in the below 6 panels (g-l). K is the number of true, simulated subtypes and B is the number of binary traits. ρ_{GE} is the gene-subtype correlation term, with $\rho_{GE} = 0$ giving non-heritable subtype statuses and $\rho_{GE} = 1$ giving perfectly heritable subtypes. h^2_{hom}, h^2_{het}, and h^2_z are the variances explained by homogeneous SNPs, heterogeneous SNPs, and main subtype effects, respectively. As in main Fig 1, solid lines have $(h^2_{hom}, h^2_{het}) = (4\%, .4\%)$, and dashed lines are reversed; in (d,e), line types define only the h^2 term not governed by the x-axis. In (a), all methods fit $K = 2$ subtypes; there is no true heterogeneity for $K = 1$, where the oracle is not defined, and for $K > 1$ and the oracle picks a true cluster at random. Generally, increasing the heterogeneous factors (h^2_{het} and h^2_z) makes subtyping easier, while increasing h^2_{hom} makes subtyping harder.
(TIF)

S3 Fig. Simulations where SNP heterogeneity only exists for some traits, for which they are only homogeneous. Left: the tested trait has no genetic heterogeneity or main subtype effect. Center: the tested trait has only a main subtype effect but no heterogeneity. Right: the full heterogeneity simulation. Linear subtype estimators (CCA and Y PC) are not trait-specific.
(TIF)

S4 Fig. Main Fig 1 with further subtyping methods. MFMR+ varies MFMR by treating the tested SNP as heterogeneous. GMM+ varies GMM by including the SNPs as traits when clustering. As expected, MFMR+ and GMM+ are miscalibrated. GMM+ often fails to converge, especially for $N \geq 10,000$ (we evaluate only the converged runs). The other methods, with low power, define subtypes as the top PC of Y or G, optionally thresholded to be binary ("G PC+disc").
(TIF)

S5 Fig. Simulations with non-Gaussian noise. Purely homogeneoues simulations, without subtypes, where the noise, ϵ, has marginal t_5 distributions. ϵ is simulated by drawing i.i.d. t_5-distributed random variables, arranging into an $N \times P$ matrix, and then right-multiplying with $\Sigma^{1/2}$, where Σ is the noise covariance matrix and is drawn as in the main simulations in main Fig 1.
(TIF)

S6 Fig. Simulations with non-linear homogeneous effects. Purely homogeneoues simulations, without subtypes, where SNPs truly have a non-linear effect. In (a-d), the SNPs are squared before use in MFMR, so that the true SNP and the utilized covariate (i.e. SNP^2) have zero correlation. In (e-h), the true SNPs are exponentiated before inclusion in MFMR, so the true SNP effect is log-linear. Results are partitioned by whether the tested traits are quantitative or binary, as well as by whether the true SNP effect is null or homogeneous.
(TIF)

S7 Fig. Simulations with continuously-varying subtypes. z is chosen to be Gaussian. Top: Effect sizes are chosen so that power roughly matches main Fig 1; it is not trivial to directly convert effect sizes from the discrete z simulations. Bottom: All heterogeneoues effect sizes are

doubled relative to top panels.
(TIF)

S8 Fig. Alternate versions of Fig 1. Top: SNP effect heterogeneity tests are applied to binary traits, not quantitative traits as in main Fig 1. Even though GMM only clusters the quantitative traits, tests for the (correlated) binary traits are miscalibrated. Middle: a 20% population prevalence binary trait is ascertained to have 50% in-sample prevalence and then tested. Bottom: population structure is added and MFMR, Oracle and GMM-PC test conditional on three genetic PCs; GMM and GMM-PC use the same subtype estimator.
(TIF)

S9 Fig. Simulations where some covariates and traits are swapped. Simulation modification where decompositions falsely treat a trait as a SNP/covariate (a,b,e,f,i) or vice versa (c,d,g,h,j). (a-d) No genetic or main subtype heterogeneity is simulated, so that the positive heterogeneity associations are unambiguously false. We test both the variable that we misplace (a,c) and the correctly place SNP/covariates and traits (b,d). (e-j) Simulations are drawn as in main text Fig 1, with $K = 2$. (e-h) Tests are shown for the misplaced trait/covariate in (e,g); for the truly homogeneous SNPs in (f,h); and for the truly heterogeneous SNPs in (i,j).
(TIF)

S10 Fig. Out-of-sample likelihood varying *K* in CONVERGE (left) and METSIM (right). Samples are split into 5 folds; parameters are fit holding one fold out; the parameters's likelihood is evaluated on the held out fold; and the process is repeated for each fold. The log-likelihoods are shown relative to the baseline likelihood of each fold at $K = 1$; this is analogous to using likelihood ratio statistics to compare a general K to the null with $K = 1$. The average across folds are shown in red, and the maximizer of K is highlighted in green.
(TIF)

S11 Fig. Metabolic subtype-specific SNP effects across all 228 NMR traits. SNP-phenotype pairs where the test for effect heterogeneity across subtypes is significant at $p = .05/81$. We test all 228 NMR-based metabolomic traits here rather than using their top PCs as in main Fig 4 and the MFMR decomposition used to learn subtypes. Per-subtype estimates and standard errors are provided in colors as in main Fig 4.
(TIF)

S12 Fig. Comparison of the $-\log_{10}(p)$-values for ordinary GWAS (x-axis) and our novel GxE GWAS (y-axis). Guide lines are drawn at $p = 5 \times 10^{-8}$, the conventional GWAS threshold. Each point is a SNP, and colors indicate which analyses were significant for the SNP. T2D, CHD, WHR and insulin are omitted because they have no genome-wide significant hits in either analysis; preT2D is omitted because the only hit is shared between both analyses. NMR PC 5 is omitted because it is badly inflated in GxE GWAS ($\lambda_{GC} = 1.83$); this trait has one hit in GWAS.
(TIF)

S13 Fig. Comparison of GREML and IID metabolic heritability estimates. Left: total IID GxEMM heritability (which adds the homogeneous and heterogeneous estimates) compared to the ordinary heritability estimated with GREML. Right: histogram of per-trait heritability increases from replacing GREML with IID GxEMM.
(TIF)

S1 Table. λ_{GC} for GWAS, MFMR GxE GWAS, and GMM GxE GWAS. GWAS means the standard regression approach conditioning on known covariates and genetic PCs. RGWAS is

our approach, which uses covariate-aware clusters (MFMR) and tests for genetic variant effect heterogeneity (Het) or globally for any genetic effect (Global). "Previous Subtyping" is like RGWAS, except using covariate-unaware clustering (GMM) and heterogeneity tests. "Inf" means our calculations suffered numerical problems, meaning that λ_{GC} is very large. We do not perform SNP tests on NMR PC 5.
(CSV)

S2 Table. Statin effect heterogeneity test for $K \in \{2, 3, 4\}$ and the 16 traits used to define clusters with MFMR. The "MFMR `droptest`" columns test using the large-effect covariate test implemented in the `rgwas` R package and described in the Methods in the main text. The "+Condition on T2D" columns run the same linear model as in `droptest`, except that T2D status is additionally included as a covariate; this analysis is performed to add confidence that the heterogeneous statin effect on glucose is not merely driven by simple confounding from T2D status.
(CSV)

Acknowledgments

We thank members of Zaitlen lab for helpful discussions and the individuals who participated in the CONVERGE and METSIM studies.

Author Contributions

Conceptualization: Andy Dahl, Noah Zaitlen.

Data curation: Andy Dahl, Na Cai, Arthur Ko, Päivi Pajukanta, Jonathan Flint.

Formal analysis: Andy Dahl.

Funding acquisition: Noah Zaitlen.

Investigation: Andy Dahl.

Methodology: Andy Dahl.

Project administration: Andy Dahl, Noah Zaitlen.

Resources: Markku Laakso, Päivi Pajukanta, Jonathan Flint, Noah Zaitlen.

Software: Andy Dahl.

Supervision: Noah Zaitlen.

Validation: Andy Dahl.

Visualization: Andy Dahl.

Writing – original draft: Andy Dahl.

Writing – review & editing: Andy Dahl, Na Cai, Arthur Ko, Markku Laakso, Päivi Pajukanta, Jonathan Flint, Noah Zaitlen.

References

1. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. Differences in breast cancer stage at diagnosis and cancer-specific survival by race and ethnicity in the United States. JAMA. 2015; 313(2):165–173. https://doi.org/10.1001/jama.2014.17322 PMID: 25585328
2. Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindström S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nature Genetics. 2017; 49 (12):1767–1778. https://doi.org/10.1038/ng.3785 PMID: 29058716

3. Dominantly Inherited Alzheimer Network, Ringman JM, Goate A, Masters CL, Cairns NJ, Danek A, et al. Genetic Heterogeneity in Alzheimer Disease and Implications for Treatment Strategies. Current Neurology and Neuroscience Reports. 2014; 14(11):429. https://doi.org/10.1007/s11910-014-0499-8
4. Jeste SS, Geschwind DH. Disentangling the heterogeneity of autism spectrum disorder through genetic findings. Nature reviews Neurology. 2014; 10(2):74–81. https://doi.org/10.1038/nrneurol.2013.278 PMID: 24468882
5. Gibson P, Tong Y, Robinson G, Thompson MC, Currle DS, Eden C, et al. Subtypes of medulloblastoma have distinct developmental origins. Nature. 2010; 468(7327):1095–1099. https://doi.org/10.1038/nature09587 PMID: 21150899
6. Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nature medicine. 2015; 21(7):730–738. https://doi.org/10.1038/nm.3897 PMID: 26121193
7. Mueller S, Engleitner T, Maresch R, Zukowska M, Lange S, Kaltenbacher T, et al. Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes. Nature. 2018; 554(7690):62–68. https://doi.org/10.1038/nature25459 PMID: 29364867
8. Flint J, Kendler KS. The genetics of major depression. Neuron. 2014; 81(3):484–503. https://doi.org/10.1016/j.neuron.2014.01.027 PMID: 24507187
9. Patel CJ, Chen R, Kodama K, Ioannidis JPA, Butte AJ. Systematic identification of interaction effects between genome- and environment-wide associations in type 2 diabetes mellitus. Human Genetics. 2013; 132(5):495–508. https://doi.org/10.1007/s00439-012-1258-z PMID: 23334806
10. Udler MS, Kim J, von Grotthuss M, Bonàs-Guarch S, Cole JB, Chiou J, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS medicine. 2018; 15(9):e1002654. https://doi.org/10.1371/journal.pmed.1002654 PMID: 30240442
11. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells. Science. 2014; 343(6175):1246980–1246980. https://doi.org/10.1126/science.1246980 PMID: 24604203
12. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Nature. 2014; 343(6175):1246949–1246949.
13. Knowles DA, Davis JR, Edgington H, Raj A, Favé MJ, Zhu X, et al. Allele-specific expression reveals interactions between genetic variation and environment. Nature Methods. 2017; 14(7):699–702. https://doi.org/10.1038/nmeth.4298 PMID: 28530654
14. Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, Richards JB, et al. Genetic interactions affecting human gene expression identified by variance association mapping. eLife. 2014; 3:e01381. https://doi.org/10.7554/eLife.01381 PMID: 24771767
15. Small KS, Todorčević M, Civelek M, Moustafa JSES, Wang X, Simon MM, et al. Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. Nature Genetics. 2018; 50(4):572–580. https://doi.org/10.1038/s41588-018-0088-x PMID: 29632379
16. Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature Genetics. 2013; 45(9):984–994. https://doi.org/10.1038/ng.2711 PMID: 23933821
17. Anttila V, Bulik-Sullivan B, Finucane HK, Bras J, Duncan L, Escott-Price V, et al. Analysis of shared heritability in common disorders of the brain. BioRxiv. 2016; p. 048991.
18. Exner DV, Dries DL, Domanski MJ, Cohn JN. Lesser response to angiotensin-converting-enzyme inhibitor therapy in black as compared with white patients with left ventricular dysfunction. New England Journal of Medicine. 2001; 344(18):1351–1357. https://doi.org/10.1056/NEJM200105033441802 PMID: 11333991
19. International Warfarin Pharmacogenetics Consortium, Klein TE, Altman RB, Eriksson N, Gage BF, Kimmel SE, et al. Estimation of the warfarin dose with clinical and pharmacogenetic data. New England Journal of Medicine. 2009; 360(8):753–764. https://doi.org/10.1056/NEJMoa0809329 PMID: 19228618
20. Mega JL, Simon T, Collet JP, Anderson JL, Antman EM, Bliden K, et al. Reduced-Function CYP2C19 Genotype and Risk of Adverse Clinical Outcomes Among Patients Treated With Clopidogrel Predominantly for PCI: A Meta-analysis. JAMA. 2010; 304(16):1821–1830. https://doi.org/10.1001/jama.2010.1543 PMID: 20978260
21. Rothwell PM, Cook NR, Gaziano JM, Price JF, Belch JFF, Roncaglioni MC, et al. Effects of aspirin on risks of vascular events and cancer according to bodyweight and dose: analysis of individual patient data from randomised trials. Lancet (London, England). 2018; 392(10145):387–399. https://doi.org/10.1016/S0140-6736(18)31133-4

22. Nicolau M, Levine AJ, Carlsson G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108(17):7265–7270. https://doi.org/10.1073/pnas.1102826108 PMID: 21482760

23. Fusi N, Lippert C, Borgwardt K, Lawrence ND, Stegle O. Detecting regulatory gene–environment interactions with unmeasured environmental factors. Bioinformatics. 2013; 29(11):1382–1389. https://doi.org/10.1093/bioinformatics/btt148 PMID: 23559640

24. Arnedo J, Svrakic DM, del Val C, Romero-Zaliz R, Hernández-Cuervo H, Consortium MGoS, et al. Uncovering the Hidden Risk Architecture of the Schizophrenias: Confirmation in Three Independent Genome-Wide Association Studies. American Journal of Psychiatry. 2015; 172(2):139–153. https://doi.org/10.1176/appi.ajp.2014.14040435 PMID: 25219520

25. Maier R, Moser G, Chen GB, Ripke S, Coryell W, Potash JB, et al. Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. The American Journal of Human Genetics. 2015; 96(2):283–294. https://doi.org/10.1016/j.ajhg.2014.12.006 PMID: 25640677

26. Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine. 2015; 7(311):311ra174–311ra174. https://doi.org/10.1126/scitranslmed.aaa9364 PMID: 26511511

27. Hinks TSC, Brown T, Lau LCK, Rupani H, Barber C, Elliott S, et al. Multidimensional endotyping in patients with severe asthma reveals inflammatory heterogeneity in matrix metalloproteinases and chitinase 3–like protein 1. Journal of Allergy and Clinical Immunology. 2016; 138(1):61–75. https://doi.org/10.1016/j.jaci.2015.11.020 PMID: 26851968

28. Wang L, Liang R, Zhou T, Zheng J, Liang BM, Zhang HP, et al. Identification and validation of asthma phenotypes in Chinese population using cluster analysis. Annals of Allergy, Asthma & Immunology. 2017; 119(4):324–332. https://doi.org/10.1016/j.anai.2017.07.016

29. Krishnan ML, Wang Z, Aljabar P, Ball G, Mirza G, Saxena A, et al. Machine learning shows association between genetic variability in PPARGand cerebral connectivity in preterm infants. Proceedings of the National Academy of Sciences. 2017; 114(52):13744–13749. https://doi.org/10.1073/pnas.1704907114

30. Nguyen QH, Lukowski SW, Chiu HS, Senabouth A, Bruxner TJC, Christ AN, et al. Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Research. 2018; 28(7):1053–1066. https://doi.org/10.1101/gr.223925.117 PMID: 29752298

31. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The Lancet Diabetes & Endocrinology. 2018. https://doi.org/10.1016/S2213-8587(18)30051-2

32. Sattar N, Preiss D, Murray HM, Welsh P, Buckley BM, de Craen AJ, et al. Statins and risk of incident diabetes: a collaborative meta-analysis of randomised statin trials. The Lancet. 2010; 375(9716):735–742. https://doi.org/10.1016/S0140-6736(09)61965-6

33. Preiss D, Seshasai SRK, Welsh P, Murphy SA, Ho JE, Waters DD, et al. Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: a meta-analysis. JAMA. 2011; 305(24):2556–2564. https://doi.org/10.1001/jama.2011.860 PMID: 21693744

34. Lotta LA, Sharp SJ, Burgess S, Perry JRB, Stewart ID, Willems SM, et al. Association Between Low-Density Lipoprotein Cholesterol–Lowering Genetic Variants and Risk of Type 2 Diabetes. JAMA. 2016; 316(13):1383–1391. https://doi.org/10.1001/jama.2016.14568 PMID: 27701660

35. Dahl A, Cai N, Flint J, Zaitlen N. GxEMM: Extending linear mixed models to general gene-environment interactions. BioRxiv. 2018.

36. Zaitlen N, Zaitlen N, Pasaniuc B, Pasaniuc B, Patterson N, Patterson N, et al. Analysis of case-control association studies with known risk variants. Bioinformatics. 2012; 28(13):1729–1737. https://doi.org/10.1093/bioinformatics/bts259 PMID: 22556366

37. Zaitlen N, Lindström S, Pasaniuc B, Cornelis M, Genovese G, Pollack S, et al. Informed Conditioning on Clinical Covariates Increases Power in Case-Control Association Studies. PLoS Genetics. 2012; 8(11):e1003032–13. https://doi.org/10.1371/journal.pgen.1003032 PMID: 23144628

38. Golan D, Rosset S. Effective Genetic-Risk Prediction Using Mixed Models. The American Journal of Human Genetics. 2014; 95(4):383–393. https://doi.org/10.1016/j.ajhg.2014.09.007 PMID: 25279982

39. Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(49):E5272–81. https://doi.org/10.1073/pnas.1419064111 PMID: 25422463

40. Weissbrod O, Flint J, Rosset S. Estimating SNP-Based Heritability and Genetic Correlation in Case-Control Studies Directly and with Summary Statistics. The American Journal of Human Genetics. 2018; 103(1):89–99. https://doi.org/10.1016/j.ajhg.2018.06.002 PMID: 29979983

41. Lindsay B, Liu J. Model Assessment Tools for a Model False World. Statistical Science. 2009; 24 (3):303–318. https://doi.org/10.1214/09-STS302

42. Dudbridge F, Fletcher O. Gene-Environment Dependence Creates Spurious Gene-Environment Interaction. The American Journal of Human Genetics. 2014; 95(3):301–307. https://doi.org/10.1016/j.ajhg.2014.07.014 PMID: 25152454

43. Consortium C. Sparse whole genome sequencing identifies two loci for major depressive disorder. Nature. 2015; 523(7562):588–591. https://doi.org/10.1038/nature14659

44. Peterson RE, Cai N, Dahl AW, Bigdeli TB, Edwards AC, Webb BT, et al. Molecular Genetic Analysis Subdivided by Adversity Exposure Suggests Etiologic Heterogeneity in Major Depression. The American journal of psychiatry. 2018; 175(6):545–554. https://doi.org/10.1176/appi.ajp.2017.17060621 PMID: 29495898

45. Laakso M, Kuusisto J, Stančáková A, Kuulasmaa T, Pajukanta P, Lusis AJ, et al. The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. Journal of Lipid Research. 2017; 58(3):481–493. https://doi.org/10.1194/jlr.O072629 PMID: 28119442

46. Kim MJ, Yu CY, Theusch E, Naidoo D, Stevens K, Kuang YL, et al. SUGP1 is a novel regulator of cholesterol metabolism. Human Molecular Genetics. 2016; 13:ddw151. https://doi.org/10.1093/hmg/ddw151

47. Aslibekyan S, Goodarzi MO, Frazier-Wood AC, Yan X, Irvin MR, Kim E, et al. Variants Identified in a GWAS Meta-Analysis for Blood Lipids Are Associated with the Lipid Response to Fenofibrate. PLoS ONE. 2012; 7(10):e48663. https://doi.org/10.1371/journal.pone.0048663 PMID: 23119086

48. Xi B, Wang C, Wu L, Zhang M, Shen Y, Zhao X, et al. Influence of Physical Inactivity on Associations Between Single Nucleotide Polymorphisms and Genetic Predisposition to Childhood Obesity. American Journal of Epidemiology. 2011; 173(11):1256–1262. https://doi.org/10.1093/aje/kwr008 PMID: 21527513

49. Qi Q, Chu AY, Kang JH, Jensen MK, Curhan GC, Pasquale LR, et al. Sugar-Sweetened Beverages and Genetic Risk of Obesity. New England Journal of Medicine. 2012; 367(15):1387–1396. https://doi.org/10.1056/NEJMoa1203039 PMID: 22998338

50. Tam CHT, Ma RCW, So WY, Wang Y, Lam VKL, Germer S, et al. Interaction Effect of Genetic Polymorphisms in Glucokinase (GCK) and Glucokinase Regulatory Protein (GCKR) on Metabolic Traits in Healthy Chinese Adults and Adolescents. Diabetes. 2009; 58(3):765–769. https://doi.org/10.2337/db08-1277 PMID: 19073768

51. Nettleton JA, McKeown NM, Kanoni S, Lemaitre RN, Hivert MF, Ngwa J, et al. Interactions of Dietary Whole-Grain Intake With Fasting Glucose- and Insulin-Related Genetic Loci in Individuals of European Descent: A meta-analysis of 14 cohort studies. Diabetes Care. 2010; 33(12):2684–2691. https://doi.org/10.2337/dc10-1150 PMID: 20693352

52. Perez-Martinez P, Corella D, Shen J, Arnett DK, Yiannakouris N, Tai ES, et al. Association between glucokinase regulatory protein (GCKR) and apolipoprotein A5 (APOA5) gene polymorphisms and triacylglycerol concentrations in fasting, postprandial, and fenofibrate-treated states. The American Journal of Clinical Nutrition. 2008; 89(1):391–399. https://doi.org/10.3945/ajcn.2008.26363 PMID: 19056598

53. Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. Exploiting Gene-Environment Interaction to Detect Genetic Associations. Human heredity. 2007; 63(2):111–119. https://doi.org/10.1159/000099183 PMID: 17283440

54. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011. https://doi.org/10.1016/j.ajhg.2010.11.011

55. Sul JH, Bilow M, Yang WY, Kostem E, Furlotte N, He D, et al. Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models. PLoS Genetics. 2016; 12(3):e1005849. https://doi.org/10.1371/journal.pgen.1005849 PMID: 26943367

56. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics. 2010; 42(7):565–569. https://doi.org/10.1038/ng.608 PMID: 20562875

57. Sladek R. The many faces of diabetes: addressing heterogeneity of a complex disease. The Lancet Diabetes & Endocrinology. 2018. https://doi.org/10.1016/S2213-8587(18)30070-6

58. Elam MB, Ginsberg HN, Lovato LC, Corson M, Largay J, Leiter LA, et al. Association of Fenofibrate Therapy With Long-term Cardiovascular Risk in Statin-Treated Patients With Type 2 Diabetes. JAMA Cardiology. 2017; 2(4):370–380. https://doi.org/10.1001/jamacardio.2016.4828 PMID: 28030716

59. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. Nature. 2017; 550(7675):239–243. https://doi.org/10.1038/nature24267 PMID: 29022581

60. Liley J, Todd JA, Wallace C. A method for identifying genetic heterogeneity within phenotypically defined disease subgroups. Nature Genetics. 2016; 49(2):310–316. https://doi.org/10.1038/ng.3751 PMID: 28024155

61. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748–752. https://doi.org/10.1038/nature08185 PMID: 19571811

62. Cortes A, Dendrou C, Motyer A, Jostins L, Vukcevic D, Dilthey A, et al. Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. BioRxiv. 2017; p. 105122.

63. Fusi N, Lippert C, Lawrence ND, Stegle O. Warped linear mixed models for the genetic analysis of transformed phenotypes. Nature communications. 2014; 5:4890. https://doi.org/10.1038/ncomms5890 PMID: 25234577

64. Robinson MR, English G, Moser G, Lloyd-Jones LR, Triplett MA, Zhu Z, et al. Genotype-covariate interaction effects and the heritability of adult body mass index. Nature Genetics. 2017; 49(8):1174–1181. https://doi.org/10.1038/ng.3912 PMID: 28692066

65. Ni G, van der Werf J, Zhou X, Hyppönen E, Wray NR, Lee SH. Genotype-covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model. BioRxiv. 2018; p. 377796.

66. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014; 507(7492):371–375. https://doi.org/10.1038/nature13138 PMID: 24646999

67. Finucane H, Reshef Y, Anttila V, Slowikowski K, Gusev A, Byrnes A, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. BioRxiv. 2017; p. 103069.

68. Zou J, Schiebinger L. AI can be sexist and racist—it's time to make it fair. Nature. 2018; 559 (7714):324–326. https://doi.org/10.1038/d41586-018-05707-8 PMID: 30018439

69. Lin Z, Yang C, Zhu Y, Duchi J, Fu Y, Wang Y, et al. Simultaneous dimension reduction and adjustment for confounding variation. Proceedings of the National Academy of Sciences. 2016; 113(51):14662–14667. https://doi.org/10.1073/pnas.1617317113

70. Abid A, Zhang MJ, Bagaria VK, Zou J. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nature communications. 2018; 9(1):2134. https://doi.org/10.1038/s41467-018-04608-8 PMID: 29849030

71. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010; 42(4):348–354. https://doi.org/10.1038/ng.548 PMID: 20208533

72. Leisch F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. Journal of Statistical Software. 2004; 11(8):1–18. https://doi.org/10.18637/jss.v011.i08

73. Prasad R, Groop L. Genetics of Type 2 Diabetes—Pitfalls and Possibilities. Genes. 2015; 6(1):87–123. https://doi.org/10.3390/genes6010087 PMID: 25774817

74. The CARDIoGRAMplusC4D Consortium, Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nature Genetics. 2015; 47(10):1121–1130. https://doi.org/10.1038/ng.3396 PMID: 26343387

75. Dahl A, Iotchkova V, Baud A, Johansson Å, Gyllensten U, Soranzo N, et al. A multiple-phenotype imputation method for genetic studies. Nature Genetics. 2016; 48(4):466–472. https://doi.org/10.1038/ng.3513 PMID: 26901065

76. Mazumder R, Hastie T, Tibshirani R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. Journal of Machine Learning Research. 2010; 11(Aug):2287–2322. PMID: 21552465