Simple multi-trait analysis identifies novel loci associated with growth and obesity measures

Xia Shen*, Xiao Wang, Zheng Ning, Yakov Tsepilov, Masoud Shirali, Generation Scotland, Blair H. Smith, Lynne J. Hocking, Sandosh Padmanabhan, Caroline Hayward, David J. Porteous, Yudi Pawitan, Chris S. Haley†, Yurii S. Aulchenko*†

Abstract: Anthropometric traits are of global clinical relevance as risk factors for a wide range of disease, including obesity[1,2]. Yet despite many hundreds of genetic variants having been associated with anthropometric measurements, these variants still explain little variation of the traits[3,4]. Joint-modeling of multiple anthropometric traits, has the potential to boost discovery power, but has not been applied to global-scale meta-analyses of genome-wide association studies (meta-GWAS). Here, we develop a simple method to perform multi-trait meta-GWAS using summary statistics reported in standard single-trait meta-GWAS and replicate the findings in an independent cohort. Using the summary statistics reported by the GIANT consortium meta-GWAS of 270,000 individuals[5], we discovered 359 novel loci significantly associated with six anthropometric traits. The "overeating gene" *GRM5* (P = 4.38×10^{-54}) was the strongest novel locus[6-8], and was independently replicated in the Generation Scotland cohort (n = 9,603, P = 4.42×10^{-3}). The novel variants had an enriched rediscovery rate in the replication cohort. Our results provide new important insights into the biological mechanisms underlying anthropometric traits and emphasize the value of combining multiple correlated phenotypes in genomic studies. Our method has general applicability and can be applied as a secondary analysis of any standard GWAS or meta-GWAS with multiple traits.

Joint-modeling of multiple traits of shared biological relevance has yet to be fully exploited in GWAS, because efficient and appropriate multivariate statistical tools are lacking. Recent efforts have indicated the potential power of jointly analyzing multiple phenotypes[9-11]. It has been noted that multi-trait statistical testing can be conducted based on standard single-trait meta-GWA summary statistics[11]. However, a general method is still needed to provide meaningful genetic effects estimates and to perform corresponding replication analysis. Here, we show that a classic multivariate analysis of variance (MANOVA) test statistic can be calculated for every genomic marker using only single-trait summary statistics, without knowing the original individual-level data (see Methods). We also demonstrate how the multi-trait genetic effect can be expressed as an additive genetic effect on a newly defined phenotype, so that the genetic effect can be interpreted and replicated in different cohorts.

We first downloaded the meta-GWAS summary statistics for six anthropometric traits: body mass index (BMI), height, weight, hip circumference, waist circumference, and waist-hip ratio, reported by the GIANT consortium[5]. In total, the summary statistics of 2,476,216 single nucleotide polymorphisms (SNPs) in common for all six single-trait meta-GWAS were passed onto subsequent analyses. Next, we estimated the correlation matrix of the six traits in the original meta-GWAS using the single-trait t-statistics of the genome-wide variants and computed the MANOVA test statistic of the six traits against each SNP (see Methods). The resulted p-values for all the variants were obtained with subsequent genomic control[12] ($\lambda = 1.001$, Supplementary Fig. 1).

The association p-values from our multi-trait meta-GWAS were compared to those from each original single-trait meta-GWAS (Fig. 1). We considered the significant SNPs located on the same chromosome and less than 500Kb apart as one locus. Among the 558 multi-trait genome-wide significant loci ($P < 5\times10^{-8}$), 99 loci overlap with at least one of the single-trait analysis results (see also Supplementary Fig. 5-6 and Supplementary Table 9). For each SNP that had a p-value less than 5×10^{-8} in any of the six single-trait meta-GWAS and in the largest meta-GWAS to-date[3,4,13], a window ± 500kb was defined. To ensure that any additional multi-trait association was in reality an extension of the single-trait locus, we excluded 100 loci located inside these windows. This resulted in 359 novel loci (Supplementary Table 1).

We ranked the newly detected SNPs according to their MANOVA p-values: the most significant four SNPs (rs669724, rs567687, rs575392 and rs12286973) were located in the intron region of the gene *GRM5* on chromosome 11. The top variant rs669724 had a p-value of 4.38×10^{-54} in a sample size of 38,800, with a minor allele frequency (MAF) 0.025 in HapMap II CEU (build 22)

(Table 1). We constructed a six-trait combined phenotype score that defined a new phenotype, **S**, based on the multiple regression of the allelic dosage of rs669724 on the six measured traits. Although the coefficients in such a multiple regression were unknown, they could be estimated from the single-trait meta-GWAS summary statistics (see Methods). We estimated the effect of rs669724 on **S** in the GIANT population as 0.0068 (s.e. = 0.0004) based on the MANOVA test statistic, which indicates that rs669724 explains 0.68% variance-covariance of the six traits (see Methods). For comparison, we estimated the phenotype score of the *FTO* locus. The top variant rs11642841 (MAF = 0.45, $P = 5.88\times10^{-56}$) at the *FTO* locus explains only 0.37% of the variance-covariance of the six traits. This indicated that the information measured by the six anthropometric traits captured by the *GRM5* rare variant rs669724 is nearly twice as much as that by the *FTO* variant rs11642841.

In the recently available Generation Scotland cohort[14], which was not a part of the GIANT analysis, we computed the same phenotype score **S** for 9,603 individuals, using the above coefficients estimated in the GIANT population. The allelic dosages of rs669724 were extracted, with a MAF of 0.003 and imputation R-square 0.77. With such a low MAF, the power of replication was limited; nevertheless, the genetic effect of rs669724 on **S** was replicated with a p-value of 4.42×10^{-3} (Table 1). Given the effect size and standard error in the GIANT population, and MAF in the Generation Scotland cohort, we estimated the 95% confidence interval of the replication p-value in the Generation Scotland cohort should be (0.0029, 0.0211), which covers our replication p-value.

Although the molecular mechanism of the multiple *GRM5* intron variants is unclear, our finding is consistent with previous reports. A large CNV (duplication) with length about 5.1Mb at the *GRM5* locus was found amongst those enriched in obese subjects[7]. The expression of *GRM5* in obese mice was significantly higher than lean mice[8]. The antagonist of *GRM5*, MTEP ($C_{11}H_8N_2S$), was shown to reduce overeating in baboons[6].

The strength of the *GRM5* multi-trait association in the GIANT meta-analysis favored replication, but we lacked sufficient power to specifically replicate other individual findings after correction for multiple testing. Nevertheless, we conducted the same replication procedure (coefficients to construct phenotype scores given in Supplementary Table 1) and obtained the replication p-values for all the newly associated SNPs. In order to examine whether our method mapped true signals, we computed the rediscovery rates (RDR)[15] of these loci in the Generation Scotland cohort (Figure 2). The RDR is defined as the proportion of SNPs replicable in the replication cohort at 5% significance threshold, given a particular p-value threshold in the discovery meta-GWAS that determines which SNPs are passed onto replication analysis. Assuming the size of each SNP effect is the same in the discovery and replication populations, we also computed the expected RDR in the Generation Scotland cohort given its sample size and allele frequencies. The results showed that our RDR across all the novel SNPs had an enrichment, not only compared to the null, but also better than the expectation when a stringent discovery threshold is applied.

Besides the *GRM5* locus, we also investigated the published biological evidence among the 25 novel meta-GWAS loci that had a p-value less than 2×10^{-16} (observed RDR larger than or equal to expected). More than half of these loci harbor candidate genes with reported relevance to obesity or obesity-associated disease (Supplementary Table 1-2). For instance, very recent evidence shows that *IRF5* (rs15498, $P = 1.90\times10^{-20}$) controls mass of adipose tissue depots and insulin sensitivity in obesity[16]. *TGFBR2* (rs6794685, $P = 3.05\times10^{-19}$) is a receptor of TGF-beta which is closely associated with BMI, obesity and type 2 diabetes[17]. *HDAC9* (rs11770723, $P = 3.38\times10^{-19}$) leads to obesity-induced body fat dysfunction and metabolic disease during high-fat feeding in mice[18], and recently, similar behaviors have been reported for *AHR* at the same locus[19].

According to the Genetic Association of Complex Diseases and Disorders (GAD) database, 252 genes at the novel loci were previously found to be associated with different types of disease, e.g. metabolic, cardiovascular, psychiatric diseases and cancer (Supplementary Table 5.4). When we conducted high-throughput functional annotation analysis using DEPICT[20] for the novel loci, but no clear enrichment of functional gene sets were found (Supplementary Table 5.2). This is consistent with results from loci identified in the single-trait meta-GWAS, for which no significant gene set enrichment was found at a false discovery rate (FDR) threshold of 5%. However, when combining the multi-trait and single-trait loci together, 7 gene sets showed FDR

< 5%, including MP:0009395 that regulates nucleated erythrocyte cell number, MP:0004810 that regulates hematopoietic stem cell number, and three GO items (GO:0040008, GO:0001558, GO:0045926) which all regulate growth (Supplementary Table 8.5).

Our analysis substantially improved the power of mapping novel variants by combining correlated traits, which is analog to combining repeated measurements of a single trait. With such power, we observed novel discoveries across different MAF values, but most of the novel variants had low MAF (Supplementary Fig. 2). Thus, we expect that more rare variants than common ones can be detected if the sample size meta-GWAS increases.

We conclude that constructing a combined phenotype score from directly measured traits adds statistical power to detect additional loci and explain missing heritability. The modified MANOVA statistic is a highly practical method that can be readily applied to any number of correlated phenotypes in large-scale association studies reliant only on summary data. Our approach holds promise for extracting further value from the ever-increasing number of large-scale meta-analyses emerging from established consortia with quantitative trait measures, including multi-omic data.

Our analysis translates each SNP-multi-trait association into a single additive effect parameter, so that replication of the genetic effect is meaningful. This is the major advantage of our method compared to previous tools[10,11]. The demonstration of equivalence between MANOVA test statistic and this additive effect is also statistically novel.

With our results, we emphasize the value of combining multiple related phenotypes in large-scale genomic studies. We expect immediate application of our method to the massive available meta-GWAS summary statistics from different global-scale consortia, which would substantially boost the discovery power and reveal more interesting biological knowledge for multiple complex traits.

LEGENDS

Figure 1: Comparison of -log$_{10}$P-values from the multi-trait and six single-trait genome-wide association meta-analyses. The dashed lines represent the genome-wide significance threshold of 5×10^{-8}.

Figure 2: Rediscovery rates of the multi-trait genome-wide association meta-analyses at different significance thresholds in the discovery population. The threshold of rediscovery was set to 0.05. The observed rediscovery rates were calculated by testing the significant SNPs that passed each discovery threshold in the replication cohort and calculating the replicated proportion. The expected rediscovery rates were estimated assuming that the effect size of each SNP is the same in both the discovery and replication populations.

Table 1: Discovery and replication summary statistics of the *GRM5* locus. β_S is the effect size on the combined phenotype score. Chr: chromosome. f: allele frequency. A: allele. The p-value for the discovery sample was obtained from the six-trait MANOVA, and the replication p-value was obtained by testing the phenotype score estimated in the discovery sample on the genotype in the replication cohort.

Supplementary Figure 1: Quantile-quantile plot of the multi-trait meta-GWAS results. The red line indicates equality, i.e. the null distribution.

Supplementary Figure 2: The multi-trait meta-GWAS results at different minor allele frequencies (MAF). A: all variants across the genome. B: novel variants.

Supplementary Figure 3: Empirical null p-value distribution of Pillai's trace statistic with the shrinkage phenotypic correlation matrix. Two traits with correlation coefficient of 0.7 were simulated. The simulated total sample size was 50,000. Genotypes of a single SNP were simulated under Hardy-Weinberg equilibrium. Prop. Overlap: proportion of sample overlap. MAF: minor allele frequency.

Supplementary Figure 4: Empirical null p-value distribution of Pillai's trace statistic with

the original phenotypic correlation matrix. Two traits with correlation coefficient of 0.7 were simulated. The simulated total sample size was 50,000. Genotypes of a single SNP were simulated under Hardy-Weinberg equilibrium. Prop. Overlap: proportion of sample overlap. MAF: minor allele frequency.

Supplementary Figure 5: Venn diagram of significant loci overlapping for meta-GWAS of multi-trait (mv), height and weight.

Supplementary Figure 6: Venn diagram of significant loci overlapping for meta-GWAS of multi-trait (mv), BMI, height and weight.

Supplementary Table 1: A list of all the novel loci mapped in the multi-trait meta-GWAS. Candidate genes with reported relevance to obesity or obesity-associated disease are highlighted in bold. Beta.S is the estimated effect in the GIANT population on the new phenotype scores, where the coefficients for constructing the new phenotypes are given in the last six columns. N is the minimum sample size among the six traits. Chr: chromosome. Freq: allele frequency. A: allele. MAF: minor allele frequency.

Supplementary Table 2: Relevance to obesity or obesity-associated disease of the candidate genes at the loci with enriched rediscovery rate ($P < 2\times10^{-16}$). Each locus is defined as a ±500kb interval centered at the most significant marker.

Supplementary Table 3: Estimated shrinkage and original phenotypic correlation matrices.

Supplementary Table 4: Average proportions of sample overlap between each pair of traits.

Supplementary Table 5: Summary of DEPICT results for the novel loci from multi-trait meta-GWAS.

Supplementary Table 6: Summary of DEPICT results for the novel loci from multi-trait

meta-GWAS at different significance thresholds.

Supplementary Table 7: Summary of DEPICT results for the significant loci from single-trait meta-GWAS at different significance thresholds.

Supplementary Table 8: Summary of DEPICT results for all the significant loci from both multi-trait and single-trait meta-GWAS at different significance thresholds.

Supplementary Table 9: List of the defined loci and the overlap across traits.

REFERENCES

1. Ng, M. *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384,** 766–781 (2014).
2. Fall, T. *et al.* The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med.* **10,** e1001474 (2013).
3. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
4. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206
5. Randall, J. C. *et al.* Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* **9,** e1003500 (2013).
6. Bisaga, A., Danysz, W. & Foltin, R. W. Antagonism of glutamatergic NMDA and mGluR5 receptors decreases consumption of food in baboon model of binge-eating disorder. *Eur Neuropsychopharmacol* **18,** 794–802 (2008).
7. Wang, K. *et al.* Large copy-number variations are enriched in cases with moderate to extreme obesity. *Diabetes* **59,** 2690–2694 (2010).
8. Brownell, A.-L. *et al.* Modulation of mGluR5 expression in brain-gut axis: Relation to obesity. *The Journal of Nuclear Medicine* (2013).
9. van der Sluis, S., Posthuma, D. & Dolan, C. V. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.* **9,** e1003235 (2013).
10. Bolormaa, S. *et al.* A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet.* **10,** e1004198 (2014).
11. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96,** 21–36 (2015).
12. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
13. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518,** 187–196 (2015).
14. Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* **42,** 689–700 (2013).
15. Ganna, A., Lee, D., Ingelsson, E. & Pawitan, Y. Rediscovery rate estimation for assessing the validation of significant findings in high-throughput studies. *Brief. Bioinformatics* (2014). doi:10.1093/bib/bbu033
16. Dalmas, E. *et al.* Irf5 deficiency in macrophages promotes beneficial adipose tissue expansion and insulin sensitivity during obesity. *Nat. Med.* (2015). doi:10.1038/nm.3829
17. Zhu, H., Kash, S. F., Drake, T. A., Sachs, A. & Lusis, A. J. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature* (2005).
18. Chatterjee, T. K. *et al.* HDAC9 knockout mice are protected from adipose tissue dysfunction and systemic metabolic disease during high-fat feeding. *Diabetes* **63,** 176–187 (2014).
19. Xu, C.-X. *et al.* Aryl hydrocarbon receptor deficiency protects mice from diet-induced adiposity and metabolic disorders through increased energy expenditure. *Int J Obes*

(Lond) (2015). doi:10.1038/ijo.2015.63

20. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6,** 5890 (2015).
21. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–75– S1–3 (2012).
22. Shen, X., Ning, Z. & Pawitan, Y. A simple regression equivalence of Pillai's trace statistic. *arXiv* 1–3
23. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23,** 1294–1296 (2007).
24. 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).

ACKNOWLEDGEMENTS

X.S. was supported by a Swedish Research Council grant (no. 2014-371). Y.S.A. was supported by RFBR grant no. 15-34-20763 and by the European Union's Seventh Framework Programme (FP7-Health) under the grant agreements no. 305280 (MIMOmics) and no. 602736 (Pain-Omics). C.S.H and M.S. were supported by the UK Medical Research Council and the UK Biotechnology and Biological Sciences Research Council (Grant BB/J002844/1). The work of Y.T. was supported by a grant from the Russian Science Foundation (RSCF, grant no. 14-14-00313).

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the UK Medical Research Council. We are grateful to all the families who took part, the general practitioners and the Scottish School of Primary Care for their help in recruiting them, and the whole Generation Scotland team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants and nurses.

We thank the Swedish Twin Registry for allowing us extract the names of a set of independent SNPs to estimate the correlation matrix of the six traits analyzed by the GIANT consortium.

AUTHOR INFORMATION

Affiliations

Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12 A, SE-17 177, Stockholm, Sweden.
Xia Shen, Zheng Ning, Yudi Pawitan

MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crew Road, EH4 2XU, Edinburgh, UK.
Xia Shen, Masoud Shirali, Caroline Hayward, Chris S. Haley

Novosibirsk State University, Pirogova 2, 630090, Novosibirsk, Russia
Yakov A. Tsepilov, Yurii S. Aulchenko

Institute of Cytology and Genetics SB RAS, Lavrentyeva ave. 10, 630090, Novosibirsk, Russia
Yakov A. Tsepilov, Yurii S. Aulchenko

Centre for Population Health Sciences, University of Edinburgh, Medical School, Teviot Place, Edinburgh, EH8 9AG, United Kingdom
Yurii S. Aulchenko

PolyOmica, De Savornin Lohmanlaan 19a, 9722, Groningen, The Netherlands.
Yurii S. Aulchenko

Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Svante Arrheniusväg 20 C, SE-10 691, Stockholm, Sweden.
Xiao Wang

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, United Kingdom.
Chris S. Haley

Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Oberschleißheim, Germany.
Yakov A. Tsepilov

Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.

David J. Porteous

Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.
Blair H. Smith, Lynne J. Hocking, Sandosh Padmanabhan, Caroline Hayward, David J. Porteous

Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK.
David J. Porteous

Division of Population Health Sciences, University of Dundee, Dundee, UK.
Blair H. Smith

Division of Applied Health Sciences, University of Aberdeen, Aberdeen, UK.
Lynne J. Hocking

Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK.
Sandosh Padmanabhan

Contributions

Initiated and coordinated the study: X.S. and Y.S.A.
Developed statistical methods: X.S., Z.N. and Y.P.
Analyzed the data: X.S.
Investigated candidate genes: X.W. and Y.T.
Contributed to replication analysis: M.S.
Contributed replication data: Generation Scotland, B.H.S., L.J.H., S.P., C.H. and D.J.P.
Supervised the study: C.S.H. and Y.S.A.
Contributed to writing: X.S., X.W., Y.T., D.J.P., C.S.H. and Y.S.A.
[†]These authors contributed equally to this work.

Competing financial interests

Dr. Yurii Aulchenko is a founder and co-owner of PolyOmica – a private research organization that specializes in consulting in statistical (gen)omics.

Corresponding authors

[*]Correspondence should be addressed to xia.shen@ki.se and y.s.aulchenko@polyomica.com.

METHODS

Anthropometric traits meta-GWAS summary statistics

We downloaded the summary statistics of six sex-stratified anthropometric traits meta-GWAS by the GIANT consortium from:

https://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

For each trait, we computed the summary statistics un-stratified by sex by meta-analyzing the effects and standard errors of the two genders. As HapMap II allele frequencies were reported in the meta-GWAS instead of pooled allele frequencies across all the cohorts, we excluded SNPs with sample size less than 30,000, for which the HapMap allele frequencies might not be representative. SNPs with missing allele frequencies were also excluded.

Generation Scotland cohort

The data were obtained from the Generation Scotland: Scottish Family Health Study (GS:SFHS)[14]. Ethical approval for the study was given by the NHS Tayside committee on research ethics (reference 05/s1401/89). Governance of the study, including public engagement, protocol development and access arrangements, was overseen by an independent advisory board, established by the Scottish government.

Individuals were genotyped with the Illumina OMNIExpress chip (706,786 SNPs). We used GenABEL version 1.7-6[23] and PLINK version 1.07[24] to exclude SNPs that had a missingness > 2% and a Hardy-Weinberg Equilibrium test $P < 10^{-6}$. Duplicate samples, individuals with gender discrepancies and those with more than 2% missing genotypes were also removed. After this quality control, the data set consisted in 9,603 individuals, genotyped for 646,127 SNPs on the 22 autosomes. Individual height, weight and waist and hip circumferences were recorded as previously described[14]. The six anthropometric phenotypes were adjusted for age and sex and inverse-Gaussian transformed. The population structure was corrected using a linear mixed model, following the procedures ibs(weight = “freq”) and polygenic() in GenABEL.

Multi-trait association modeling

For k phenotypes, where k is often much less than the sample size n, the association between the group of k phenotypes and a biallelic marker $\mathbf{g}$ can be expressed as a multivariate regression

$$\mathbf{Y}_{n\times k} = \mathbf{1}_{n\times 1}\boldsymbol{\mu}'_{k\times 1} + \mathbf{g}_{n\times 1}\boldsymbol{\beta}'_{k\times 1} + \mathbf{e}_{n\times k} \tag{1}$$

which can be tested via MANOVA for the null hypothesis

$$H_0 : \boldsymbol{\beta} = \mathbf{0} \tag{2}$$

As in most GWA analyses, here, each phenotypic vector in $\mathbf{Y}$ is adjusted for other covariates and inverse-Gaussian transformed to be standard-normal distributed. The estimates in the vector $\hat{\boldsymbol{\beta}}$ are known from GWA summary statistics. Below, we show how a MANOVA test statistic can be obtained without knowing the original data. 421

Calculating the multi-trait association test statistic

We derive the multi-trait association test statistic based on the summary statistics from each single univariate meta-GWAS. Assuming the genotype of each individual i follows Hardy-Weinberg equilibrium (HWE), and the marker minor allele frequency (MAF) is f, we have

$$E[\mathbf{y}] = \mathbf{1}\boldsymbol{\mu}' + E[\mathbf{g}]\boldsymbol{\beta}' = \mathbf{1}\boldsymbol{\mu}' + 2f\boldsymbol{\beta}' = \mathbf{0} \tag{3}$$

Thus, $\hat{\boldsymbol{\mu}} = -2f\hat{\boldsymbol{\beta}}$. The residual variance-covariance matrix of (1) is

$$\mathbf{E} = (\mathbf{Y} - \mathbf{1}\hat{\boldsymbol{\mu}}' - \mathbf{g}\hat{\boldsymbol{\beta}}')'(\mathbf{Y} - \mathbf{1}\hat{\boldsymbol{\mu}}' - \mathbf{g}\hat{\boldsymbol{\beta}}') \tag{4}$$

The corresponding residual variance-covariance matrix of the null model is

$$\mathbf{E}_0 = (\mathbf{Y} - E[\mathbf{Y}])'(\mathbf{Y} - E[\mathbf{Y}]) \tag{5}$$

After some simple math, we have

$$\mathbf{E}_0 = n\mathbf{R} \tag{6}$$

422

where $\mathbf{R}_{k\times k}$ is the correlation matrix of the k phenotypes. Similarly,

$$\mathbf{E} = \mathbf{E}_0 - \mathbf{H} \tag{7}$$

where

$$\mathbf{H} = 2nf(1-f)\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}' \tag{8}$$

$\mathbf{H}$ is the model variance-covariance matrix captured by the marker. Anolog of the univariate ANOVA F-test, let λ_j ($j = 1, \ldots, k$) be the eigenvalues solving $\det(\mathbf{H} - \lambda\mathbf{E}) = 0$. "Pillai's trace" can be constructed as

$$V = tr(\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}) = \sum_{i=1}^{k} \frac{\lambda_i}{1+\lambda_i} \tag{9}$$

and the corresponding F-statistic is

$$\frac{V/k}{(1-V)/(n-k-1)} \sim F_{k,n-k-1} \tag{10}$$

When n is large, the F-statistic is approximately $\chi^2(k)$-distributed.

Shrinkage estimate of the phenotypic correlation matrix

If $\mathbf{R}$ is unknown, it can also be estimated using the single-trait GWAS summary statistics. If T_j and $T_{j'}$ are the t-statistics of phenotypes $\mathbf{y}_j$ and $\mathbf{y}_{j'}$ against a particular variant in the GWAS, then we have the correlation coefficient $R_{j,j'} = cor(T_j, T_{j'})$ if $\mathbf{y}_j$ and $\mathbf{y}_{j'}$ are measured in the same population[11]. So that $\mathbf{R}$ can be estimated by selecting a large number of independent variants from the GWA summary statistics and calculating their correlation matrix. In most meta-GWAS, as in our study, some traits are not measured in all cohorts, namely, the individuals in single-trait GWAS partially overlap. In such case, assuming the individuals measured in $\mathbf{y}_j$ is a subset of $\mathbf{y}_{j'}$, let $\mathbf{g}$ and $\mathbf{x}$ be the genotypes of the overlapping and non-overlapping individuals. We have

$$\begin{aligned} cor(T_j, T_{j'}) &= \frac{\mathbf{g}'cov(\mathbf{y}_j, \mathbf{y}_{j'})(\mathbf{g}', \mathbf{x}')'}{\sqrt{\mathbf{g}'\mathbf{g}\sigma_j^2}\sqrt{(\mathbf{g}', \mathbf{x}')(\mathbf{g}', \mathbf{x}')'\sigma_{j'}^2}} && (11) \\ &= \frac{R_{j,j'}\sigma_j\sigma_{j'}\mathbf{g}'(\mathbf{I}, \mathbf{0})(\mathbf{g}', \mathbf{x}')'}{\sqrt{\mathbf{g}'\mathbf{g}\sigma_j^2}\sqrt{(\mathbf{g}', \mathbf{x}')(\mathbf{g}', \mathbf{x}')'\sigma_{j'}^2}} && (12) \\ &= \sqrt{\theta_{j,j'}}R_{j,j'} && (13) \end{aligned}$$

where $\theta_{j,j'}$ is the proportion of overlapping individuals between traits $\mathbf{y}_j$ and $\mathbf{y}_{j'}$. Therefore, an unbiased estimate of $R_{j,j'}$ can be obtained by calculating

$$R_{j,j'} = \sqrt{\theta_{j,j'}^{-1}}cor(T_j, T_{j'}) \tag{14}$$

However, simulations in **Supplementary Figure 4** indicate that directly using such an unbiased estimate of $\mathbf{R}$ inflates Pillai's trace statistic and generates more false positives than expected. We therefore use $cor(T_j, T_{j'})$ as a shrinkage estimate of $R_{j,j'}$ in Pillai's trace statistic, although such shrinkage results in underpowered testing **Supplementary Figure 3**.

In order to obtain a set of independent SNPs in the GIANT population, we performed LD-pruning using PLINK option `--indep-pairwise 50 5 0.1` on the genotype data of 644,556 SNPs typed in 9,741 unrelated individuals from the Swedish Twin Registry, which is a cohort included in the GIANT population. T-statistics of the resulted 49,036 independent SNPs were used to estimate the shrinkage phenotypic correlation matrix.

Constructing the new phenotype score

We construct a new phenotype score as a linear combination of the original six phenotypes, via the following multiple regression model,

$$\mathbf{g} = \mathbf{1}a + \mathbf{Y}\mathbf{b} + \boldsymbol{\epsilon} \tag{15}$$

Now we show how the coefficients estimates in eq. (15) can also be obtained without knowing the original data. First of all, we derive the coefficients estimates of each *swapped* GWA simple regression model,

$$\mathbf{g} = \mathbf{1}a_j^* + \mathbf{y}_j b_j^* + \boldsymbol{\epsilon}_j^* \tag{16}$$

From the summary statistics, we know the estimates of the following GWA simple regression model,

$$\mathbf{y}_j = \mathbf{1}\mu_j + \mathbf{g}\beta_j + \mathbf{e}_j \tag{17}$$

i.e.

$$\hat{\beta}_j = \frac{\mathbf{g}'\mathbf{y}_j - \bar{\mathbf{g}}\bar{\mathbf{y}}_j}{\mathbf{g}'\mathbf{g} - \bar{\mathbf{g}}^2} = \frac{\mathbf{g}'\mathbf{y}_j}{2f(1-f)} \tag{18}$$

since $\mathbf{y}_j \sim N(\mathbf{0}, \mathbf{I})$, and Hardy-Weinberg equilibrium (HWE) is assumed. We also have

$$\hat{b}_j^* = \frac{\mathbf{y}_j'\mathbf{g} - \bar{\mathbf{g}}\bar{\mathbf{y}}_j}{\mathbf{y}_j'\mathbf{y}_j - \bar{\mathbf{y}}_j^2} = \mathbf{y}_j'\mathbf{g} \tag{19}$$

As $\mathbf{g}'\mathbf{y}_j = \mathbf{y}_j'\mathbf{g}$, we have

$$\hat{b}_j^* = 2f(1-f)\hat{\beta}_j \tag{20}$$

Thereafter, the estimates $\hat{\mathbf{b}}$ in eq. (15) can be calculated as

$$\hat{\mathbf{b}} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{D}\hat{\mathbf{b}}^* \tag{21}$$

where $\mathbf{D}$ is a diagonal matrix with the j-th element $\mathbf{y}_j'\mathbf{y}_j$[21]. Again, since $\mathbf{y}_j \sim N(\mathbf{0}, \mathbf{I})$, we obtain

$$\hat{\mathbf{b}} = 2f(1-f)\mathbf{R}^{-1}\hat{\mathbf{b}}^* \tag{22}$$

So that a new phenotype score can be defined as

$$\mathbf{S} = \mathbf{Y}\hat{\mathbf{b}} \tag{23}$$

Estimating the genetic effect on the new phenotype score

In a replication cohort, the genetic effect of each SNP on the new phenotype score $\mathbf{S}$ can be tested via simple regression of $\mathbf{S}$ on the allelic dosages $\mathbf{g}$,

$$\mathbf{S} = \mu_s + \mathbf{g}\beta_s + \mathbf{e}_s \tag{24}$$

Interestingly, without knowing the original data, we can obtain the estimate of β_s in the meta-GWAS population. We showed elsewhere[22] that $\hat{\beta}_s$ always equals to Pillai's trace V in eq. (9). Translating the MANOVA p-value back to a 1 d.f. χ^2 statistic C, we can compute the standard error of $\hat{\beta}_s$ in the meta-GWAS population as $VC^{-1/2}$.

Also, we showed that $\beta_s = V = R^2$, where R^2 is the coefficient of determination of both regressions (15) and (24) in the meta-GWAS population. Therefore, Pillai's trace V directly represents the proportion of the variance of $\mathbf{S}$ explained by the SNP.

Genomic control

In a large sample, the null distribution of our test statistic for the six traits is asymptotically chi-square with 6 degrees of freedom. We estimated the genomic inflation factor λ as the ratio of the observed median chi-square value across the genome to its expectation 5.348. The estimated λ = 1.001, thus the chi-square values were divided by the estimated λ, and the corresponding genomic-controlled p-values were reported.

Locus definition

For each trait and multi-trait (MV) meta-GWAS, significant loci were defined by collapsing adjacent markers. We selected the SNPs with p-value $< 5\times10^{-8}$ and checked if they were located on the same chromosome and less than 500 Kb apart - we considered these SNPs as one locus and used the most significant SNP to represent the locus. This resulted in 656 significant loci across all meta-GWAS, including 558 loci from MV GWAS, 158 for height, and 50, 30, 11, 7, 5 for weight, BMI, waist circumference (WC), hip circumference (HIP) and waist-hip-ratio (WHR), respectively. For two meta-GWAS, if the top variants at a locus are less than 500 Kb apart, the locus is considered overlapping between the two meta-GWAS (see also Supplementary Table 9 and Supplementary Fig. 5-6).

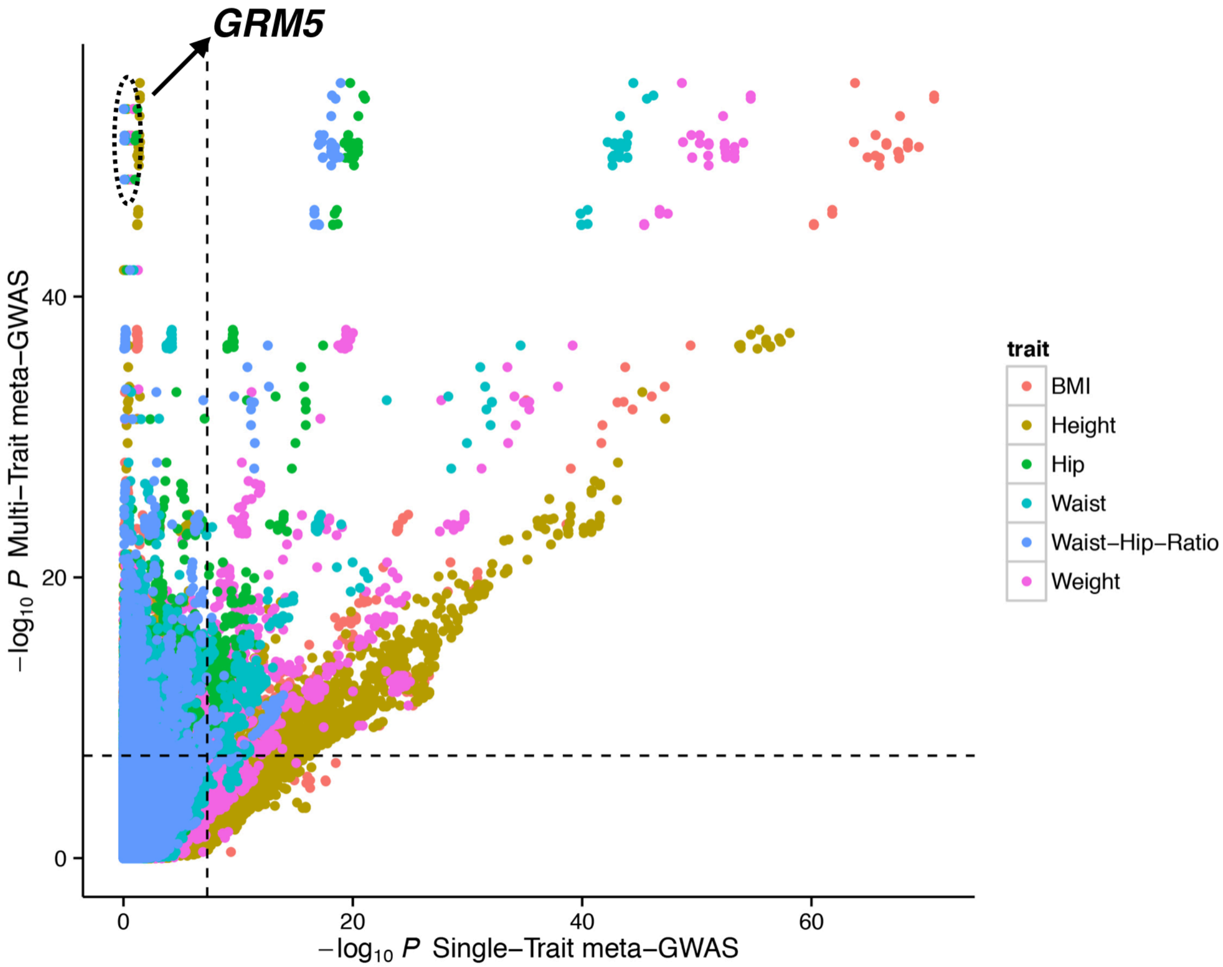
Functional annotation

We conducted high-throughput functional annotation of the novel discoveries. For prioritizing genes in associated regions, gene set enrichment and tissue/cell type enrichment, we used the DEPICT software[20]. We first analyzed the loci that were found using single-trait GWAS only, then we analyzed all the loci that were found using the MV approach, thereafter we analyzed all these loci together (Supplementary Table 5-8). In each step, we applied two thresholds: $P < 1\times10^{-9}$ and $P < 1\times10^{-16}$, in order to compare the results at different RDR (see main text).

Availability

The developed multi-trait GWA method is implemented and freely available in the MultiSummary() procedure of the R package **MultiABEL** (The **GenABEL** project packages URL: https://r-forge.r-project.org/R/?group_id=505).

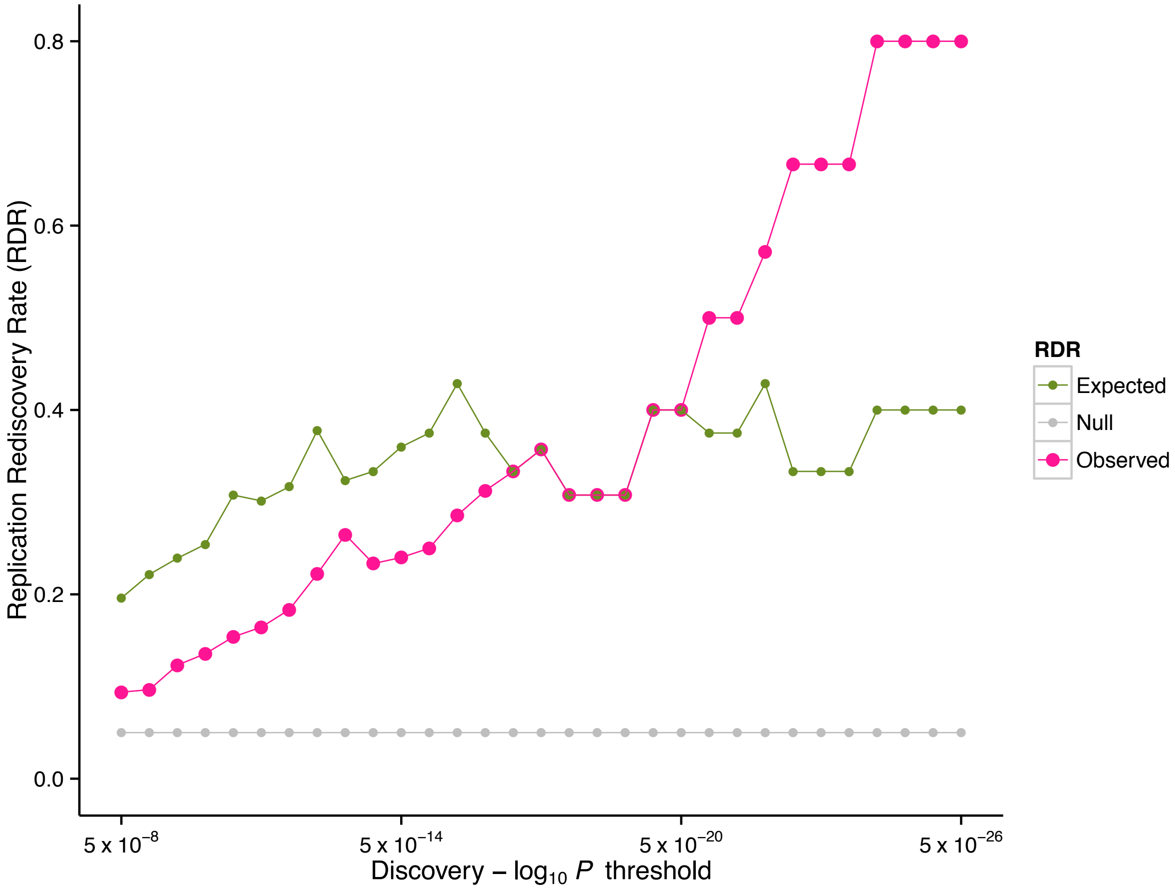
464 Figure 1



465

466

467 Figure 2



468

469

Table 1

Leading Variant	Chr	Candidate Gene	A1	A2	Population	*N*	f(A1)	β_s (s.e.)	*P*
rs669724	11	*GRM5*	A	G	GIANT (discovery)	38,800	0.025	0.0068 (0.0004)	4.38E-54
					Generation Scotland (replication)	9,603	0.003	0.0044 (0.0016)	4.42E-03