Fine-scale population structure in the UK Biobank

DOI:
10.1093/hmg/ddaa157

Document Version
Accepted author manuscript

Fine-scale population structure in the UK Biobank: implications for genome-wide association studies

James P Cook[1], Anubha Mahajan[2,] and Andrew P Morris[1,2, 3*]

[1]Department of Biostatistics, University of Liverpool, Liverpool, UK. [2]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [3]Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, University of Manchester, Manchester, UK.

[*]Prof Andrew Morris

Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Division of Musculoskeletal and Dermatological Sciences, The University of Manchester, AV Hill Building, Upper Brook Street, Manchester M13 9LJ, United Kingdom

E-mail: andrew.morris-5@manchester.ac.uk

Tel: 0161 306 8732

Fax: N/A

ABSTRACT

The UK Biobank is a prospective study of more than 500,000 participants that has aggregated data from questionnaires, physical measures, biomarkers, imaging and follow-up for a wide range of health-related outcomes, together with genome-wide genotyping supplemented with high-density imputation. Previous studies have highlighted fine-scale population structure in the UK on a North-West to South-East cline, but the impact of unmeasured geographical confounding on genome-wide association studies (GWAS) of complex human traits in the UK Biobank has not been investigated. We considered 368,325 white British individuals from the UK Biobank, and performed GWAS of their birth location. We demonstrate that widely used approaches to adjust for population structure, including principal components

analysis and mixed modelling with a random effect for a genetic relationship matrix, cannot fully account for the fine-scale geographical confounding in the UK Biobank. We observe significant genetic correlation of birth location with a range of lifestyle-related traits, including body-mass index and fat mass, hypertension, and lung function, even after adjustment for population structure. Variants driving associations with birth location are also strongly associated with many of these lifestyle-related traits after correction for population structure, indicating that there could be environmental factors that are confounded with geography that have not been adequately accounted for. Our findings highlight the need for caution in the interpretation of lifestyle-related trait GWAS in UK Biobank, particularly in loci demonstrating strong residual association with birth location.

INTRODUCTION

The United Kingdom (UK) is located off the north-western coast of the European mainland, and incorporates Great Britain, Northern Ireland, and many smaller islands (including the Hebrides, Shetlands and Orkneys). Previous studies have highlighted that population structure within the UK is rather limited, but it occurs at fine-scale on North-South and East-West clines[1,2]. Analyses undertaken using genome-wide genotyping data from the People of the British Isles collection identified genetic clusters that are highly localised, separating the Orkney Islands, Scotland and Northern England, Central and Southern England, and Wales[3]. Such fine-scale structure can lead to false positive signals in genome-wide association studies (GWAS) of traits with characteristics that vary between regions, if not adequately accounted for in the analysis[4].

Multivariate statistical techniques, such as principal components analysis (PCA), have been widely used in population genetics to visualise genotype differences between individuals in few dimensions via eigenvalue decomposition of a genetic relationship matrix (GRM). Axes of genetic variation, derived from PCA, can be used to adjust for population structure by their inclusion as covariates in a generalised linear regression model[5]. An alternative, widely-used approach to account for population structure is to adjust for the genetic correlation between individuals, as measured by the GRM, which can be included as a random effect in a generalised linear mixed model[6-13]. However, the performance of these approaches to adequately account for unmeasured confounding due to fine-scale structure in large, population-based samples has not been evaluated.

The UK Biobank is a very large and detailed prospective study of more than 500,000 participants aged 40-69 years when recruited between 2006 and 2010[14]. The study has aggregated (and continues to collect) extensive information from participants, including data

from questionnaires, physical measures, biomarkers, imaging and follow-up for a wide range of health-related outcomes (including linkage to primary care and disease-specific registers). Genome-wide genotyping data, typed on the Affymetrix UK Biobank or BiLEVE arrays, have been centrally called and quality control assessed by the UK Biobank Analysis Team[15], and imputed up to reference panels from the 1000 Genomes Project[16], UK10K Project[17] and Haplotype Reference Consortium[18]. PCA was also centrally performed by the UK Biobank Analysis Team to generate axes of genetic variation that can be used to identify participants of similar ancestry and to control for population structure[15].

In this investigation, we first assess the extent of fine-scale population structure in a subset of unrelated white British participants from the UK Biobank using demographic data of reported birth location. We then evaluate the impact of population structure on GWAS of complex human traits in the UK Biobank by considering genetic correlation with birth location and inflation in genome-wide association summary statistics. Finally, we consider locus-specific impact of residual confounding of birth location with complex human traits and demonstrate the effect on association signals of alternative approaches to account for population structure.

RESULTS

Extent of population structure in the UK Biobank. To assess the extent of fine-scale population structure in the UK Biobank, we considered a subset of unrelated white British participants based on self-reported ethnicity and centrally derived axes of genetic variation (**Materials and Methods, Supplementary Figure 1**). We then interrogated demographic data of reported birth location, for which UK postcodes had been converted to Easting and Northing Cartesian coordinates, which we refer to as "Eastings" and "Northings", respectively (**Supplementary Figure 2**). We excluded individuals with missing birth location and those from the pilot study at the Stockport recruitment centre for which the Cartesian coordinates were incorrect. For the remaining 368,325 individuals, we then tested for association of Eastings and Northings with 8,806,946 well-imputed variants with minor allele frequency (MAF) >0.5% in a linear regression model, including only genotyping array as a covariate, as implemented in SNPTESTv2.5.2[19]. To account for population structure, we then considered inclusion of: (i) the first ten (or twenty) centrally derived axes of genetic variation from PCA as covariates as implemented in SNPTESTv2.5.2[19]; or (ii) a random effect for the GRM as implemented in BOLT-LMMv2.3[13] (**Materials and Methods, Supplementary Figure 3**).

As expected, there was substantial genome-wide inflation in the association with Northings and Eastings, assessed via the LD-score regression intercept[20], with no correction for population structure (λ_N=7.817 and λ_E=6.638). Substantial inflation was also observed after adjustment for ten axes of genetic variation as covariates (λ_N=3.871 and λ_E=1.912), which was not diminished by inclusion of an additional ten axes (**Supplementary Figure 4**). The inflation was reduced after inclusion of a random effect for the GRM, but considerable fine-scale population structure remained unaccounted for: λ_N=1.651 and λ_E=1.431 (**Figure 1**).

We observed no difference in inflation between directly genotyped (λ_N=1.650 and λ_E=1.436) and imputed variants (λ_N=1.648 and λ_E=1.428). For this mixed model analysis, we observed strong negative genetic correlation between Northings and Eastings from LD-score regression[21] (r_G=-0.660, p=2.1x10^{-11}), confirming previous reports of the North-West to South-East cline in UK population structure[1]. The residual association with Northings was more pronounced than for Eastings (**Figure 1**). A total of 74 attained genome-wide significant evidence of association (p<5x10^{-8}) with Northings after inclusion of a random effect for the GRM (**Table 1**). The strongest association signals mapped to/near *TLR10-TLR1* (rs4543123, p_N=5.3x10^{-56}, p_E=2.0x10^{-12}) and *LCT* (rs1849, p_N=1.7x10^{-17}, p_E=2.0x10^{-12}), both of which have been previously reported as confounded with UK population structure[1] (**Supplementary Figures 5 and 6**). The toll-like receptor family of genes encode proteins that play a key role in the innate immune system, such that population structure could have arisen through historical geographical differences in exposure to pathogens. The *LCT* gene encodes the lactase protein that allows lactose tolerance to persist into adulthood and has been subject to positive selection after the domestication of cattle across Europe[22].

Impact of population structure on GWAS of complex human traits in the UK Biobank. We next sought to assess the impact of fine-scale UK population structure on GWAS of complex human traits in the UK Biobank. To do this, we first used LD-score regression[21] to assess the genome-wide genetic correlation between Northings and Eastings (after inclusion of a random effect for the GRM), and selected traits available in the UK Biobank. We utilised published association summary statistics available from LD-Hub[23], obtained from analysis of 337,199 unrelated white British individuals in a linear regression model with adjustment for the first ten centrally derived axes of genetic variation from PCA as covariates (**Materials and Methods**). Of the 597 traits reported in LD-Hub, we excluded those that were not

directly related to health outcomes, lifestyle and/or anthropometric measures (such as current employment, diseases of family members, education, and medication). For the remaining 268 traits, we observed significant correlation with Northings (p<0.00019, Bonferroni correction) for 41 traits (**Supplementary Table 1**), most of which were broadly related to lifestyle factors, even after adjustment for population structure. A more northerly (and westerly) birth location was genetically correlated with increased body-mass index (BMI) and fat mass, alcohol consumption, hypertension and smoking, and with decreased lung function (**Figure 2**), suggesting that association signals reported for these traits in UK Biobank could be partially driven by residual confounding with geography that has not been adequately accounted for in the analysis.

To further investigate the consequences of this residual confounding, we considered BMI and forced vital capacity (FVC, a measure of lung function) as representative of lifestyle-related traits that are genetically correlated with birth location (**Materials and Methods**). For both traits, the LD-score regression intercepts obtained from 368,325 unrelated white British individuals after inclusion of a random effect for the GRM in the linear regression model indicated evidence of residual population structure that has not been accounted for in the analysis: λ_{BMI}=1.155 and λ_{FVC}=1.099. In contrast, when we considered asthma, a disease that is characterised by poor lung function, but that did not demonstrate significant genetic correlation with birth location (p=0.70 for Northings), the impact of residual population structure was much less pronounced: λ_{ASTHMA}=1.059.

Previous studies have highlighted that genome-wide inflation in GWAS of complex human traits after inclusion of a random effect for the GRM in the linear regression model can reflect environmental factors that are confounded with geography[24], which can better be controlled for through adjustment for axes of genetic variation from PCA[25]. We hypothesised that we could account for this residual confounding of BMI and FVC by adjusting for ten

axes of genetic variation, Northings and Eastings as covariates in the linear mixed model, in addition to a random effect for the GRM (**Materials and Methods**). We demonstrated that these additional adjustments only marginally reduced the LD-score regression intercept for both traits: λ_{BMI}=1.140 and λ_{FVC}=1.095 (**Figure 3**). The same adjustments also had no impact on the LD-score regression intercept for asthma: λ_{ASTHMA}=1.057. Genome-wide, adjustment for Northings and Eastings as covariates in the linear regression model, in addition to the ten axes of genetic variation, did not have a major impact on allelic effect estimates and association *p*-values (**Supplementary Figure 7**).

We also investigated the possibility that current residence would better reflect ongoing exposure to environmental factors that are confounded with geography than would birth location. We repeated our analyses of BMI, FVC and asthma, after adjustment for Northings and Eastings derived from current residence postcode, but this did not substantially reduce the genome-wide inflation, compared with birth location, for any of these traits: λ_{BMI}=1.150, λ_{FVC}=1.096 and λ_{ASTHMA}=1.054.

Locus-specific impact of residual confounding of birth location with lifestyle-related traits in the UK Biobank. We next investigated the locus-specific impact of residual confounding of birth location with the 41 (mostly lifestyle-related) traits that were genetically correlated with Northings. To do this, we considered the 74 loci attaining genome-wide significant evidence of association ($p<5x10^{-8}$) with Northings after inclusion of a random effect for the GRM. We first dissected association signals for Northings at each locus through approximate conditional analyses implemented in GCTA[26], making use of 5,000 randomly selected white British individuals from UK Biobank as a reference for linkage disequilibrium. We identified 115 distinct association signals attaining locus-wide significance ($p<10^{-5}$) for Northings, including six mapping to the major histocompatibility complex (**Supplementary**

Table 2). Index variants for 59 (51.3%) of these signals were of low frequency (MAF<5%), which would be expected to have arisen due to more recent mutation events, and hence be more likely to be confounded with geography (**Supplementary Figure 8**).

For each distinct association signal, we then identified "high-confidence" variants accounting for at least 5% of the posterior probability of driving confounding with Northings (**Materials and Methods**). We interrogated each high-confidence variant for association with the 41 traits demonstrating significant genetic correlation with Northings in the UK Biobank. We utilised published association summary statistics available from PhenoScanner[27,28], obtained from analysis of 337,199 unrelated white British individuals in a linear regression model with adjustment for the first ten centrally derived axes of genetic variation from PCA as covariates (**Materials and Methods**). High-confidence variants driving distinct signals for Northings at five loci were associated, at genome-wide significance, with at least one trait (**Supplementary Table 3**).

At the *LCT* locus, two high-confidence variants (rs182549 and rs309137, together accounting for 66.5% of the posterior probability of driving the confounding with Northings) were associated (at genome-wide significance) with 16 of the 41 traits that were genetically correlated with birth location. The Northing increasing alleles at the two variants were associated with increased BMI and multiple measures of fat mass, and with decreased lung function (FVC and forced expiratory volume in 1-second), which are concordant with the direction of the genetic correlation with birth location. Adjustment for ten axes of genetic variation, Northings and Eastings as covariates in the linear mixed model, in addition to a random effect for the GRM, reduced the strength of association with these traits by an order of magnitude across the locus, reflecting correction for residual confounding with birth location (**Supplementary Figure 9**). There was a more noticeable impact on the association with BMI, where the estimated allelic effect of rs182549 increased four-fold after adjustment

(**Supplementary Table 4**). These results indicate the potential bias in allelic effect estimates on complex traits that could arise with inadequate correction for population structure in UK Biobank.

At the major histocompatibility complex (MHC), where population structure reflects strong selective pressure of infectious diseases in recent human history[22], one high-confidence variant (rs9268556, 13.2% posterior probability of driving the confounding with Northings) was associated (at genome-wide significance) with FVC. In contrast to the signal at the *LCT* locus, the Northing increasing allele was associated with increased FVC, which is discordant with the direction of the genetic correlation with birth location. Consequently, adjustment for ten axes of genetic variation, Northings and Eastings as covariates in the linear mixed model, in addition to a random effect for the GRM, did not noticeably reduce the strength of association with lung function at this locus (**Supplementary Table 4**).

DISCUSSION

We have demonstrated that fine-scale population structure in the UK Biobank cannot be fully accounted for through adjustment for centrally derived axes of genetic variation or inclusion of a random effect for the GRM. There was substantial inflation in genome-wide association with Northing and Easting cartesian coordinates that were derived from birth location, even after inclusion of a random effect for the GRM in the linear regression model. The inflation was greater for Northings than for Eastings, which may reflect greater variation in latitude than longitude for participants in the UK Biobank. Investigations previously undertaken with GWAS from the People of the British Isles collection indicated that major clusters separate from North to South, which could reflect major historical events in the peopling of the British Isles[3]. These results are consistent with observations across the wider European continent, where the first axis of genetic variation, which correlates with North-South geography,

explains more variability in allele frequencies than the second axis, which correlates with East-West geography[29]. Bivariate analysis of Northings and Eastings, taking account of the correlation between longitude/latitude of birth location, might provide additional insight into population structure. However, further methodological development and software is required to implement bivariate linear mixed models that can accommodate the scale of GWAS in the UK Biobank.

After correction for population structure, we have observed significant genetic correlation of Northings with 41 traits, most of which are related to lifestyle, including BMI and fat mass, alcohol consumption, hypertension, and smoking and lung function. LD-score regression intercepts for two exemplar lifestyle-related traits, BMI and FVC, indicated evidence of residual population structure that has not been accounted for by the inclusion of a random effect for the GRM in the linear regression model. Such inflation could reflect environmental factors that are confounded with geography, such as diet and smoking habits, which can better be controlled for through adjustment for axes of genetic variation. However, adjustment for ten axes of genetic variation, in addition to Eastings and Northings derived from birth location or current residence, did not substantially reduce the inflation. These results suggest that simple modelling of birth location (or current residence) and/or axes of genetic variation does not capture the full extent of geographical confounding with these environmental influences on lifestyle-related traits. More complex models, for example that allow for non-linear relationships with geography, may offer improved control for confounding with environmental risk factors, but cannot be easily accommodated in computationally efficient software that can be applied to the scale of GWAS in the UK Biobank.

We identified 74 loci that demonstrated significant residual association with Northings after inclusion of a random effect for the GRM in the linear regression model,

which map to/near genes that have been subject to selection, including *LCT* and the MHC region. High-confidence variants driving distinct residual associations for Northings were also strongly associated with many of the lifestyle-related traits that are genetically correlated with birth location, even after correction for population structure. These signals could, therefore, represent false positive associations with lifestyle-related traits that are driven by confounding with geography. At signals for which the high-confidence variant was also associated with the lifestyle-related trait in the direction predicted by the genetic correlation, such as for BMI and FVC at the *LCT* locus, additional adjustment for axes of genetic variation and birth location as covariates reduced the strength of the association. In contrast, when the association with the lifestyle-related trait was in the opposite direction to that predicted by the genetic correlation, for example for FVC in the MHC region, adjustment for axes of genetic variation and birth location as covariates had no impact on the signal. Thus, whilst adjustment for axes of genetic variation and birth location, in addition to a random effect for the GRM, did not substantially reduce the inflation in association with lifestyle-related traits genome-wide, we did observe locus-specific differences in the impact of this correction that reflect varying levels of confounding with geography.

In conclusion, our findings highlight the need for caution in the interpretation of GWAS of lifestyle-related health outcomes in UK Biobank, particularly in loci demonstrating strong residual association with birth location, even after adjustment for population structure. To minimise the impact of population structure on these traits at loci that are most strongly confounded with geography, we recommend adjusting for axes of genetic variation and birth location, in addition to a random effect for the GRM in a regression model. Where substantial residual inflation in the genome-wide association remains, for example an LD-score intercept of the order of 1.1 or more, we suggest careful consideration of potential environmental risk factors for the trait that could have more complex confounding with geography than can be

accommodated by simple linear relationships with birth location (or current residence). UK Biobank has collected extensive questionnaire data on diet, smoking, alcohol consumption and exercise, and these potential confounders can be included directly as covariates in a regression model, without any assumptions about their correlation with geography. Further studies are warranted in other large-scale biobanks, particularly in less homogenous populations where the impact of geographical confounding of allele frequencies on complex trait GWAS may be even more pronounced.

MATERIALS AND METHODS

Selection of participants from UK Biobank. We utilised the subset of "white British" individuals identified centrally by the UK Biobank Analysis Team[15], based on self-reported ethnicity from the assessment centre questionnaire and axes of genetic variation from principal components analysis. We then utilised the relatedness report generated by the UK Biobank Analysis Team[15] to retain the maximal set of unrelated participants, which corresponded to a maximum kinship coefficient of 0.0884.

We interrogated demographic data of reported birth location, for which UK postcodes had been converted to Easting and Northing Cartesian coordinates, rounded to the nearest 500m, relative to an origin in the South West of the British Isles (**Supplementary Figure 2**). We excluded individuals with missing birth location and those from the pilot study at the Stockport recruitment centre for which the Cartesian coordinates were incorrect. For some sensitivity analyses, we also considered Easting and Northing Cartesian coordinates derived from current residence postcode.

Genome-wide association analyses with Cartesian coordinates of birth location in UK Biobank. The UK Biobank Central Analysis Team performed initial quality control of

variants, and imputation up to reference panels from the 1000 Genomes Project[16], UK10K Project[17] and Haplotype Reference Consortium[18]. We considered the subset of variants that were imputed to the Haplotype Reference Consortium, excluding those with MAF <0.5% and/or imputation quality info score <0.5. For each variant passing quality control, we tested for association with Northings and Eastings, separately, in a linear regression model, using the genotype dosage from imputation, and including only genotyping array (UK Biobank or UK BiLEVE) as a covariate, as implemented in SNPTESTv2.5.2[19]. We used two approaches to account for population structure. First, we included ten centrally derived axes of genetic variation from PCA, in addition to genotyping array, as covariates as implemented in SNPTESTv2.5.2[19]. We also performed sensitivity analyses including twenty centrally derived axes of genetic variation from PCA. Second, we included a random effect for the GRM, in addition to a fixed effect for genotyping array, as implemented in BOLT-LMMv2.3[13]. We followed recommendations from the BOLT-LMM UK Biobank analysis pipeline: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-510009. The GRM was constructed from directly genotyped variants that passed initial quality control from the UK Biobank Central Analysis Team. BOLT-LMM performs "leave-one-chromosome-out" analysis: variants from the chromosome being tested for association are excluded from the GRM to avoid proximal contamination[13].

For each analysis, to assess inflation in association signals due to residual population structure that was not accounted for in the analysis, we calculated the intercept from LD-score regression[20], using a subset of approximately one million variants for which European ancestry LD-scores were available. For some sensitivity analyses, we separated directly genotyped and imputed variants, and calculated the LD-score intercept for each set.

Genetic correlation of birth location with complex human traits in the UK Biobank. We used LD-score regression[21] to assess the genome-wide genetic correlation between birth location and selected traits available in the UK Biobank. We utilised published association summary statistics available from LD-Hub[23], obtained from analysis of 337,199 unrelated white British participants passing central quality control in a generalised linear regression model with adjustment for sex and the first ten centrally derived axes of genetic variation from PCA as covariates, as implemented in Hail. Phenotypes were derived and harmonised with PHESANT[30], and association analyses were restricted to variants with MAF >0.1%, exact Hardy-Weinberg equilibrium (HWE) $p>10^{-10}$, and imputation quality info score >0.8. Full details of the quality control, phenotype derivation and association analyses can be found at: https://github.com/Nealelab/UK_Biobank_GWAS/tree/master/imputed-v2-gwas#sample-and-variant-qc.

Of the 597 traits reported from UK Biobank in LD-Hub, we excluded those that were not directly related to health outcomes, lifestyle and/or anthropometric measures (such as current employment, diseases of family members, education, and medication). For each the remaining 268 traits, we calculated the genetic correlation with Northings and Eastings using association summary statistics after adjusting for population structure by including a random effect for the GRM as implemented in BOLT-LMMv2.3[13], as described above. The LD-score regression analysis was restricted to a subset of approximately one million variants for which European ancestry LD-scores were available, and which overlapped with those reported for birth location and the complex trait. We extracted the genetic correlation, corresponding standard error and *p*-value. We defined significant genetic correlation by $p<0.00019$, which corresponded to a Bonferroni correction for 268 traits.

Genome-wide association analyses with body mass index (BMI), forced vital capacity (FVC) and asthma in UK Biobank. We performed inverse rank normalisation of BMI and FVC (best measure). For each variant passing quality control, we tested for association with each trait (after transformation), separately, in a linear regression model, using the genotype dosage from imputation, and including genotyping array (UK Biobank or UK BiLEVE) as a covariate and a random effect for the GRM as implemented in BOLT-LMMv2.3[13]. We repeated each of these analyses by including: (i) the first ten centrally derived axes of genetic variation from PCA as additional covariates; and (ii) the first ten centrally derived axes of genetic variation from PCA, Northings and Eastings as additional covariates in the linear regression model. We also repeated our analyses, adjusting for Northings and Eastings derived from current location postcode, instead of birth location postcode, in addition the first ten centrally derived axes of genetic variation from PCA. For each trait, for each analysis, we calculated the intercept from LD-score regression[20], using a subset of approximately one million variants for which European ancestry LD-scores were available.

Dissection and fine-mapping of association signals with birth location in UK Biobank. We considered each locus attaining genome-wide significant evidence of association ($p<5x10^{-8}$) with Northings. Within each locus, we utilised the "--cojo-slct" option in GCTA[26] to identify index variants representing distinct association signals attaining locus-wide significance ($p<10^{-5}$), based on: (i) association summary statistics for Northings after adjusting for population structure by including a random effect for the GRM as implemented in BOLT-LMMv2.3[13], as described above; and (ii) 5,000 randomly selected white British participants included in our association analyses as a reference for LD in the UK population. For each locus with more than one index variant, we next dissected each distinct association signal. For each index variant, we obtained the corresponding conditional association signal

by utilising the "--cojo-cond" option in GCTA[26] by adjusting for all other index variants at the locus.

Within each locus, for each distinct signal, we first approximated the Bayes' factor[31] in favour of association with Northings of each variant on the basis of summary statistics after adjusting for population structure by including a random effect for the GRM as implemented in BOLT-LMMv2.3[13], as described above. We utilised summary statistics from unconditional analysis for loci with a single signal, and GCTA conditional analysis for loci with multiple distinct signals. Specifically, the Bayes' factor for the jth variant at the ith distinct association signal is approximated by

$$\Lambda_{ij} = \exp\left[\frac{b_{ij}^2}{2v_{ij}}\right],$$

where b_{ij} and v_{ij} are the allelic effect on Northings and the corresponding variance, respectively. We then calculated the posterior probability that the jth variant is driving the ith distinct association, given by

$$\pi_{ij} = \frac{\Lambda_{ij}}{\sum_k \Lambda_{ik}},$$

where the summation is over all variants across the locus. We defined "high-confidence" variants as having posterior probability of at least 5% of driving distinct association signals for birth location.

Association of high-confidence variants with complex human traits in UK Biobank. We extracted association summary statistics for each high-confidence variant for each trait

attaining significant genetic correlation with Northings in UK Biobank using PhenoScanner[27,28]. Association summary statistics were obtained from analysis of 337,199 unrelated white British participants passing central quality control in a generalised linear regression model with adjustment for sex and the first ten centrally derived axes of genetic variation from PCA as covariates, as implemented in Hail. Phenotypes were derived and harmonised with PHESANT[30], and association analyses were restricted to variants with MAF >0.1%, exact HWE $p>10^{-10}$, and imputation quality info score >0.8.

ACKNOWLEDGEMENTS

This study has been conducted using the UK Biobank resource (project number 15390).

CONFLICT OF INTEREST STATEMENT

As of January 2020, A.M. is an employee of Genentech, and a holder of Roche stock.

REFERENCES

1. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678.
2. O'Dushlaine, C.T., Morris, D., Moskvina, V., Kirov, G., International Schizophrenia Consortium, Gill, M., Corvin, A., Wilson, J.F., and Cavalleri, G.L. (2010). Population structure and genome-wide patterns of variation in Ireland and Britain. *Eur. J. Hum. Genet.* **18**, 1248-1254.
3. Leslie S., Winney B., Hellenthal G., Davison D., Boumertit A., Day T., Hutnik K., Royrvik E.C., Cunliffe B., Wellcome Trust Case Control Consortium 2, et al. (2015). The fine-scale genetic structure of the British population. *Nature* **519**, 309-314.
4. Heath, S.C., Gut, I.G., Brennan, P., McKay, J.D., Bencko, V., Fabianova, E., Foretova, L., Georges, M., Janout, V., Kabesch, M., et al. (2008). Investigation of the fine structure of European populations with applications to disease association studies. *Eur. J. Hum. Genet.* **16**, 1413-1429.
5. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904-909.

6. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348-354.
7. Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355-360.
8. Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459-463.
9. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833-835.
10. Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nat. Methods* **9**, 525-526.
11. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821-824.
12. Svishcheva, G.R., Axenovich, T.I., Belonogova, N.M., van Duijn, C.M., and Aulchenko, Y.S. (2012). Rapid variance components-based method for whole-genome association analysis. *Nat. Genet.* **44**, 1166-1170.
13. Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284-290.

14. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779.

15. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209.

16. 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526**, 68-74.

17. UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90.

18. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, HM., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279-1283.

19. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499-511.

20. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291-295.

21. Bulik-Sullivan, B.K., Finucane, H.K., Antilla, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen Consortium, Psychiatric Genetics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., Perry, J.R.B, et al. (2015). An atlas of genetic correlation across human diseases and traits. *Nat. Genet.* **47**, 1236-1241.

22. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. *Science* **312**, 1614-1620.

23. Zheng, J., Erzurumluoglu, A.M., Elsworth, B.L., Kemp, J.P., Howe, L., Haycock, P.C., Hemani, G., Tansey, K., Laurin, C., Early Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, et al. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279.

24. Haworth, S., Mitchell, R., Corbin, L., Wade, K.H., Dudding, T., Budu-Aggrey, A., Carslake, D., Hemani, G., Paternoster, L., Smith. G.D., et al. (2019). Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333.

25. Zhang, Y., and Pan, W. (2015). Principal component regression and linear mixed model in association analysis of structured samples: competitors or complements? *Genet. Epidemiol.* **39**, 149-155.

26. Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369-375.

27. Staley, J.R., Blackshaw, J., Kamat, M.A., Ellis, S., Surendran, P., Sun, B.B., Paul, D.S., Freitag, D., Burgess, S., Danesh, J., et al. (2016). PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207-3209.

28. Kamat, M.A., Blackshaw, J.A., Young, R., Surendran, P., Burgess, S., Danesh, J., Butterworth, A.S., and Staley, J.R. (2019). PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851-4853.

29. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. *Nature* **456**, 98-101.

30. Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G., and Tilling, K. (2018). PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29-35.

31. Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

LEGENDS TO FIGURES

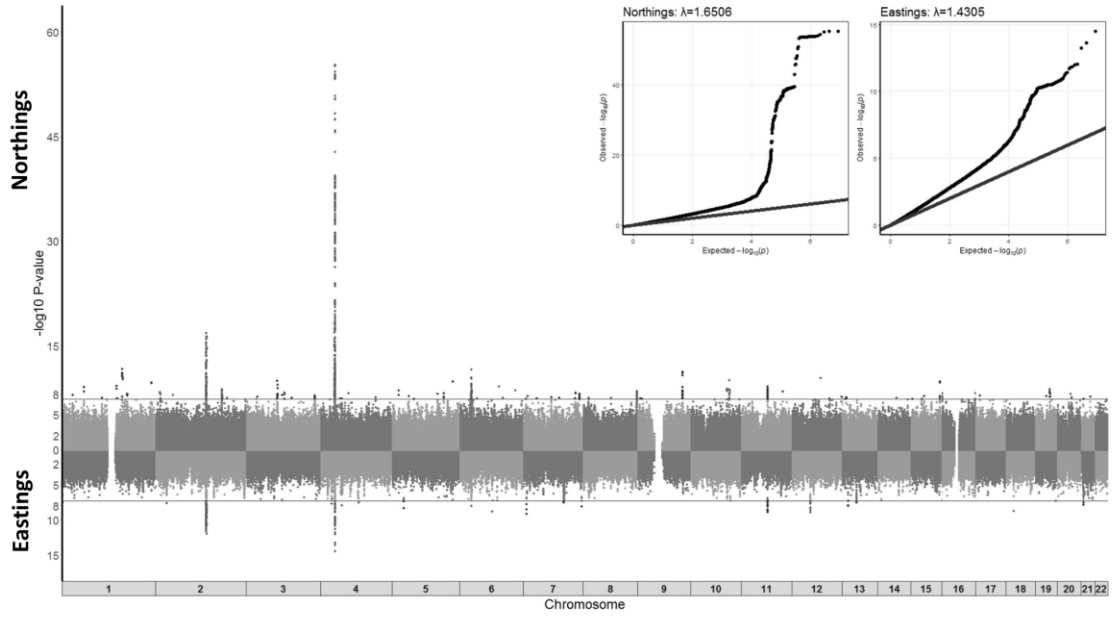

Figure 1. Miami plot and quantile-quantile plots for association with Northing and Easting Cartesian co-ordinates for birth location of unrelated white British individuals from the UK Biobank after correction for population structure. Association analyses are performed with inclusion of a random effect for the GRM in a linear mixed model. Inflation factors (λ) assessed via LD-score regression intercept. The genome-wide significance threshold (p<5x10^{-8}) is indicated by the horizontal lines.

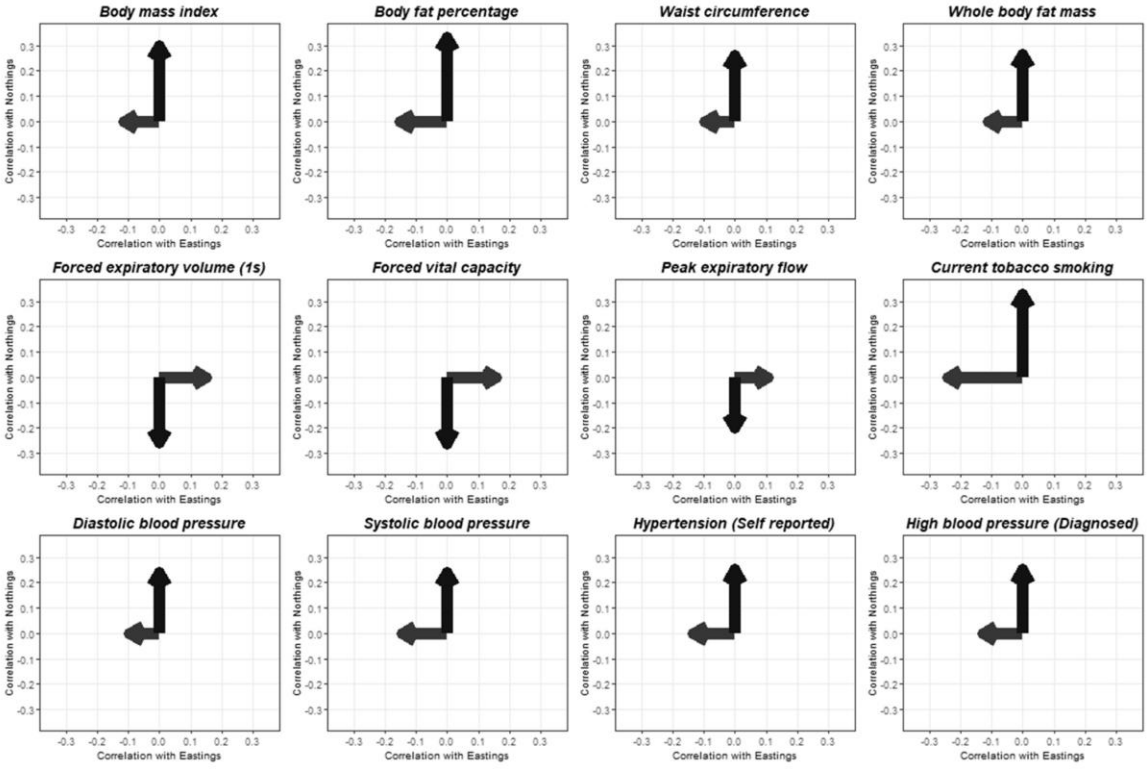

Figure 2. Genetic correlation of lifestyle related traits with Northing and Easting Cartesian co-ordinates for birth location of unrelated white British individuals from the UK Biobank. We selected twelve traits as representative of obesity and fat distribution, lung function and smoking, and blood pressure and hypertension. The tips of the arrows correspond to the genetic correlation of the trait with Northings and Eastings. A more

northerly (and westerly) birth location was genetically correlated with increased body-mass index and fat mass, hypertension and smoking, and with decreased lung function.

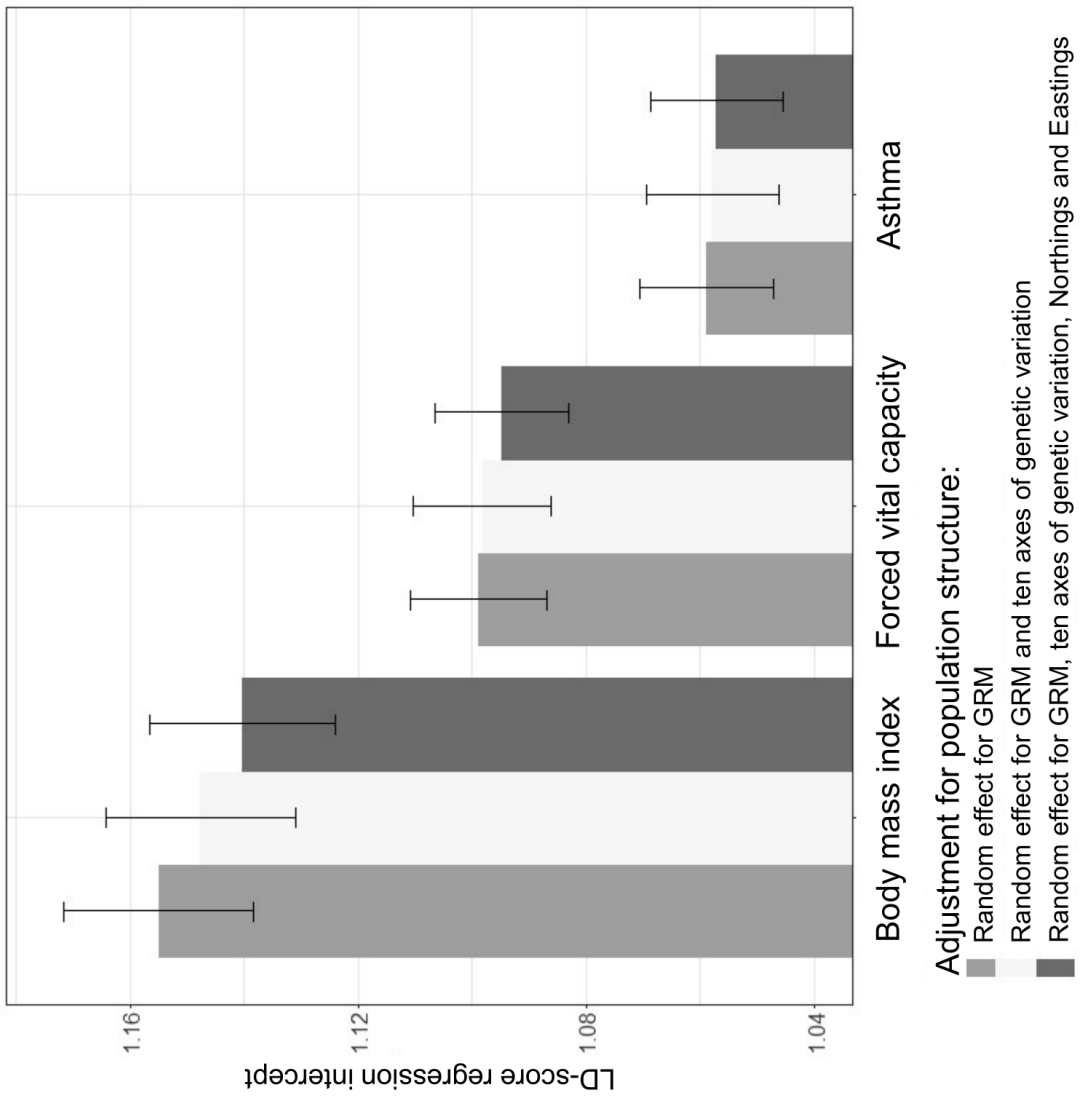

Figure 3. LD-score regression intercepts for body mass index, forced vital capacity and asthma, obtained for unrelated white British individuals from the UK Biobank after correction for population structure through inclusion of a random effect for the GRM in a linear mixed model, with and without adjustment for ten axes of genetic variation, and Northing and Easting Cartesian coordinates. The height of each bar represents the LD-score intercept, and the error bars define the 95% confidence interval.

Table 1. Loci attaining genome-wide significant association ($p<5x10^{-8}$) with Northing Cartesian coordinates for birth location of unrelated white British individuals from the UK Biobank after correction for population structure through inclusion of a random effect for the GRM in a linear mixed model.

Locus	Lead variant	Chr	Position (bp, b37)	Mixed model *p*-value	
				Northings	Eastings
YTHDF2	rs183909650	1	29,059,553	$1.2x10^{-8}$	0.75
MYSM1-JUN	rs138938527	1	59,196,687	$9.3x10^{-10}$	0.35
Intergenic	rs11184903	1	106,972,375	$3.4x10^{-8}$	0.011
POLR3C	rs141333427	1	145,599,750	$6.1x10^{-10}$	0.55
GJA5-GJA8	rs76713613	1	147,307,666	$4.1x10^{-8}$	0.49
FCRLB	rs6700369	1	161,691,586	$2.6x10^{-12}$	0.24
KIAA0040	rs2861158	1	175,135,829	$9.4x10^{-9}$	0.063
CHRM3	rs142495445	1	239,889,366	$2.1x10^{-10}$	0.023
LPIN1	rs869162	2	12,017,846	$5.3x10^{-9}$	0.82
PRKCE-EPAS1	rs72795609	2	46,458,369	$1.3x10^{-8}$	0.56
LCT	rs182549	2	136,616,754	$1.7x10^{-17}$	$2.0x10^{-12}$
PDE11A	rs75313639	2	178,613,409	$1.9x10^{-9}$	0.083
STAT4	rs17768109	2	191,920,448	$3.2x10^{-8}$	0.34
Intergenic	rs138897148	3	30,414,016	$2.4x10^{-8}$	0.53
GBE1-LINC00971	rs75932529	3	82,986,685	$1.2x10^{-10}$	0.76
Intergenic	rs189809665	3	95,199,900	$1.3x10^{-8}$	0.97
Intergenic	rs191077151	3	102,612,554	$5.6x10^{-9}$	0.61
ILDR1	rs147965995	3	121,719,991	$3.5x10^{-8}$	0.78
YEATS2	rs166398	3	183,446,977	$1.5x10^{-8}$	0.53
TLR10-TLR1	rs4543123	4	38,792,524	$5.3x10^{-56}$	$4.1x10^{-11}$
Intergenic	rs562248335	4	53,210,826	$3.6x10^{-8}$	0.55
AASDH	rs10010544	4	57,202,676	$3.7x10^{-8}$	0.23
PARM1-LINC02483	rs142147881	4	76,126,259	$6.9x10^{-9}$	0.87
SLC10A7-POU4F2	rs138838211	4	147,525,948	$2.7x10^{-8}$	0.49

LINC02100-RF00017	rs144164550	5	18,838,724	$2.7x10^{-9}$	0.96
Intergenic	rs11738948	5	44,999,799	$1.7x10^{-8}$	0.31
PART1	rs3887175	5	59,790,456	$4.7x10^{-8}$	0.050
CSNK1G3	rs2897789	5	122,948,316	$8.6x10^{-9}$	0.047
SMIM33	rs13181561	5	138,850,905	$7.6x10^{-9}$	0.00059
RP11-541P9.3	rs185543831	5	162,606,973	$1.5x10^{-10}$	0.29
MHC region	rs67850286	6	32,207,912	$2.9x10^{-12}$	0.27
ANKRD66-MEP1A	rs9463249	6	46,747,864	$3.7x10^{-8}$	0.60
RN7SKP211	rs77691922	6	106,389,862	$8.1x10^{-10}$	0.49
LINC02534	rs527638681	6	116,060,967	$3.3x10^{-8}$	0.0034
ZC3H12D-PPIL4	rs183211514	6	149,809,239	$2.7x10^{-9}$	0.14
ZNF316	rs9640029	7	6,685,123	$4.5x10^{-8}$	0.038
THSD7A-TMEM106B	rs12699279	7	11,886,719	$1.7x10^{-8}$	0.27
LOC401324	rs7807834	7	35,355,874	$4.2x10^{-8}$	0.95
GTF2IRD2	rs145191771	7	74,285,390	$2.5x10^{-8}$	0.96
AC002451.1-DYNC1I1	rs73241153	7	95,321,530	$3.2x10^{-8}$	0.013
KLRG2-CLEC2L	rs6467860	7	139,190,020	$5.8x10^{-9}$	0.38
GIMAP4	rs6969418	7	150,262,584	$8.4x10^{-9}$	0.97
TUSC3	rs12543949	8	15,309,705	$3.2x10^{-8}$	0.55
FGF20	rs2467176	8	16,692,687	$4.9x10^{-8}$	0.078
LY96	rs11466004	8	74,941,275	$3.3x10^{-8}$	0.37
JRK-PSCA	rs2920288	8	143,753,289	$3.7x10^{-9}$	0.92
KDM4C	rs140546025	9	6,765,320	$3.9x10^{-8}$	0.71
Intergenic	rs72712132	9	12,297,698	$4.7x10^{-8}$	0.57
TLR4	rs4986790	9	120,475,302	$5.8x10^{-12}$	0.014
PIK3AP1	rs12572544	10	98,509,591	$1.8x10^{-9}$	0.57
TMEM180	rs74908306	10	104,233,229	$9.6x10^{-11}$	0.055
NADSYN1-KRTAP5-7	rs11234014	11	71,232,811	$7.2x10^{-10}$	$1.1x10^{-7}$
C11orf53	rs7934982	11	111,149,632	$4.6x10^{-9}$	0.078
CSRP2	rs10746288	12	77,261,098	$4.6x10^{-11}$	0.37
MYBPC1	rs10860766	12	102,064,667	$4.3x10^{-8}$	0.12
GALNT9	rs117340324	12	132,683,244	$2.4x10^{-8}$	0.25
LINC00417-ANKRD20A9P	rs9552508	13	19,354,675	$2.6x10^{-8}$	0.65
FLT3	rs35263155	13	28,652,999	$4.5x10^{-8}$	0.76
LINC00398-LINC00545	rs73165012	13	31,388,774	$2.9x10^{-8}$	0.60
DCAF5	rs143797681	14	69,498,428	$3.1x10^{-8}$	0.69
PWRN2	rs544128806	15	24,494,412	$4.7x10^{-8}$	0.26
SECISBP2L-COPS2	rs62009762	15	49,389,757	$4.5x10^{-8}$	0.0053
PRTG-NEDD4	rs150276168	15	56,067,643	$1.8x10^{-8}$	0.34
LINC00923	rs72752662	15	98,370,408	$1.6x10^{-10}$	0.25
IFT140	rs117492052	16	1,655,759	$6.7x10^{-9}$	0.19
MC1R	rs1805007	16	89,986,117	$6.8x10^{-9}$	0.46
LINC00670	rs149081560	17	12,503,649	$1.5x10^{-8}$	0.42
ZNF536	rs149713626	19	30,817,216	$1.9x10^{-8}$	0.37
SELENOV	rs8102247	19	40,008,118	$2.1x10^{-9}$	0.46
LTBP4-NUMBL	rs2604861	19	41,150,922	$9.4x10^{-9}$	0.0026
VSTM2L	rs6013469	20	36,558,660	$9.7x10^{-9}$	0.58
MAFB	rs6102086	20	39,281,690	$2.6x10^{-8}$	0.0046

LINC01549	rs193267476	21	18,710,258	$4.7x10^{-8}$	0.31
RUNX1	rs564634064	21	36,479,812	$2.4x10^{-8}$	0.18

ABBREVIATIONS

BMI: body mass index

FVC: forced vital capacity

GRM: genetic relationship matrix

GWAS: genome-wide association study

MAF: minor allele frequency

PCA: principal components analysis