

methySCOPA and META-methySCOPA: software for the analysis and aggregation of epigenome-wide association studies of multiple correlated phenotypes

Harmen Draisma^{1,2}, Jun Liu³, Igor Pupko^{1,2}, Ayşe Demirkan^{1,3,4}, Zhanna Balkhiyarova^{1,2}, Andrew P Morris^{5,6,7}, Reedik Magi⁵, Matthias Wielscher⁸, Saqib Hassan⁹, Cornelia M. van Duijn^{3,10,11}, Sylvain Sebert¹², Marjo-Riitta Jarvelin^{8,12}, Marika Kaakinen^{1,2,13}, Inga Prokopenko^{1,2}

¹Department of Clinical & Experimental Medicine, School of Biosciences & Medicine, University of Surrey, Guildford, United Kingdom

²Section of Genomics of Common Disease, Department of Medicine, Imperial College London, London, United Kingdom

³Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands

⁴Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands

⁵Estonian Genome Center, University of Tartu, Tartu, Estonia

⁶Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK

⁷Department of Biostatistics, University of Liverpool, Liverpool, UK

⁸Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom

⁹Department of Medicine, Imperial College London, United Kingdom

¹⁰Netherlands Metabolomics Centre, Leiden University, Leiden, the Netherlands

¹¹Nuffield Department of Population Health, Oxford University, Oxford, United Kingdom

¹²Center for Life Course Health Research, University of Oulu, Oulu, Finland

¹³Centre for Pharmacology and Therapeutics, Department of Medicine, Imperial College London, London, United Kingdom

Abstract

Background: Multi-phenotype genome-wide association studies (MP-GWAS) of correlated traits have greater power to detect genotype–phenotype associations than single-trait GWAS. However, no multi-phenotype analysis method exists for epigenome-wide association studies (EWAS).

Results: We extended the SCOPA approach developed by us to “methyISCOPA” software in C++ by ‘reversely’ regressing DNA hyper/hypo-methylation information on a linear combination of phenotypes. We evaluated two models of association between DNA methylation and fasting glucose (FG) and insulin (FI) levels: Model 1, including FG, FI, and three measured potential confounders (body mass index [BMI], fasting serum triglyceride levels [TG], and waist/hip ratio [WHR]), and Model 2, including FG and FI corrected for the effects of BMI, TG, and WHR. Both models were additionally corrected for participant sex and smoking status (current/ever/never). We meta-analyzed the cohort-specific MP-EWAS results with our novel software META-methyISCOPA, mapped genomic locations to CGCh37/hg19, and adopted $P < 1 \times 10^{-7}$ to denote epigenome-wide significance. We used the Illumina Infinium HumanMethylation450K BeadChip array data from the Northern Finland Birth Cohorts (NFBC) 1966/1986. We quality-controlled the data, regressed out the effects of measured potential confounders, and normalized the methylation signal intensity and FI data. The MP-EWAS included data for 643/457 individuals from NFBC1966 and NFBC1986, respectively (total N=1,100).

In Model 1, we detected epigenome-wide significant association in the MP-EWAS meta-analysis at cg13708645 (chr12:121,974,305; $P = 1.2 \times 10^{-8}$) within *KDM2B* gene. Single-trait effects within *KDM2B* were on FI, BMI, and WHR. Model with effect on BMI and WHR showed the strongest association at this locus, while effect on FI in single-phenotype analysis was driven by the effect of adiposity. In Model 2, the strongest association was at cg05063096 (chr3:143,689,810; $P = 2.3 \times 10^{-7}$) annotated to *C3orf58* with strongest effect on FI in single-trait analysis and multi-phenotype effect on FI and WHI within Model 1.

We characterized the effects of established EWAS loci for diabetes and its risk factors and detected suggestive ($p < 0.01$) associations at six markers including *PHGDH*, *TXNIP*, *SLC7A11*, *CPT1A*, *MYO5C* and *ABCG1*, through the dissection of the multi-phenotype effects in Model 1.

Conclusions: We implemented MP-EWAS in methylSCOPA and demonstrated its enhanced power over single-trait EWAS for correlated phenotypes in large-scale data.

Keywords: Methylation, Epigenome-wide association study, Multivariate analysis, Reverse regression, Correlation, Meta-analysis

Background

Multi-phenotype genome-wide association studies (MP-GWAS) of correlated traits are more powerful, give better precision of estimates, and provide enhanced biological insight, i.e. suggestion of potential pleiotropic effects, as compared to single-phenotype GWAS¹⁻⁶. We have previously developed an MP-GWAS method using the “reverse regression” approach in which allele dosage is regressed on a linear combination of phenotypes, implemented into the software tool SCOPA and meta-analysis tool METASCOPA⁷. However, no multi-phenotype epigenome-wide association study (MP-EWAS) method exists, although EWAS have recently gained increased attention due to advances in technology and thus lowered costs of measuring epigenetic regulation.

DNA methylation is a type of epigenetic regulation and is most widely used within EWAS. Methylation refers to the attachment of methylation groups to the DNA molecule. Methylation of CpG islands within a gene’s promoter usually implies that that the gene is not transcribed. DNA methylation is tissue-specific, reversible, and inheritable. Usually, the cytosine copies on both strands are either methylated or unmethylated.

In correlated traits, there is a considerable, although incomplete, overlap between the measures of glucose homeostasis and type 2 diabetes (T2D). For instance, the genetic correlation between fasting glucose levels (FG) and fasting insulin levels (FI) estimated by cross-trait LD Score regression is 0.31, and the genetic correlations between FG and T2D and between FI and T2D are 0.58 and 0.48, respectively⁸. The study of glycaemic traits in healthy individuals can provide insights about the pathophysiology of T2D, and the (epi)genetic study on these phenotypes can inform on the molecular mechanisms leading to T2D – also those influenced by individual's lifestyle and environmental exposures, as they have been shown to leave a mark on the individual's epigenome⁹. One of the advantages of studying glycaemic traits, rather than T2D, is that sample sizes can be much larger, as they are independent of T2D prevalence. Indeed, genome-wide methylation in blood has been associated with body mass index, T2D and measures of glucose metabolism^{10–12}. However, no study has previously aimed at unravelling the epigenetics of these traits by taking into account their correlations with each other, most likely due to the lack of appropriate methodology.

Our aims in the current work were two-fold. First, we aimed to extend the reverse regression approach for methylation data and implement it in a software tool. We addressed this aim by developing methylSCOPA (Software for CORrelated Phenotype Analysis with methylation data), which is the SCOPA extension for DNA methylation data. methylSCOPA association summary statistics can also be aggregated across EWAS through fixed-effects meta-analysis, implemented in META-methylSCOPA, which is the META-SCOPA extension for MP-EWAS meta-analysis. Analogous to META-SCOPA, this enables application of reverse regression in large-scale international consortia efforts where, for instance, ethical concerns and legal restrictions preclude joint analysis of individual-level genome-wide DNA methylation and phenotype data from different studies. Second, we aimed to test the method for epigenetic effects on FG & FI variability. We report one novel methylation probe associated with FG and FI from these analyses

and dissect the multi-phenotype epigenetic effects at 11 established methylation marks for metabolic traits.

Implementation

Reverse regression model of multiple correlated phenotypes

methyISCOPA extends the SCOPA analysis framework⁷ to the analysis of DNA methylation data. Specifically, this means that methyISCOPA allows for methylation (instead of genotype as in SCOPA) data as input, and that it analyses methylation data analogous to the way in which SCOPA analyses genotype dosage data.

DNA methylation assays in tissue samples return at any given site an average methylation percentage for a mixture of cells. These percent methylation values are continuous and range from 0 to 100¹³. In methyISCOPA we model these percent methylation values as a function of the observed phenotypes using linear reverse regression, analogous to how SCOPA models the genotype at a single-nucleotide polymorphism (SNP) as a function of the observed phenotypes. Therefore, analogous to expression (1) from Mägi *et al.*⁷, considering a sample of unrelated individuals with J phenotypes denoted by y_1, y_2, \dots, y_J , in methyISCOPA we model the DNA methylation value $Methylation_i$ at a particular probe for individual i as

$$Methylation_i = \alpha + \sum_j \beta_j y_{ij} + \epsilon_i$$

In this expression, y_{ij} denotes the phenotypic value for individual i , phenotype j ; β_j denotes the effect of the j th phenotype on the degree of DNA hyper/hypomethylation at the probe (analogous to the effect β_j of the j th phenotype on genotype at the SNP under consideration in the SCOPA model); and $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is the residual variance. We recommend that covariates relevant to the multi-phenotype effects should be included in the model; otherwise, confounding factors should be regressed out of the phenotypes and resulting residuals should be used instead.

For further dissection of epigenome-wide significant ($p < 1 \times 10^{-7}$) multiple-phenotype association signals, the analysis of different phenotype combinations is enabled. We assess the model fit within each phenotype combination through the use of the Bayesian Information Criterion (BIC), with the smallest value indicating the best fit.

Meta-analysis of multiple EWAS of the same set of correlated phenotypes is enabled through the application of the method for the synthesis of regression slopes¹⁴, similarly to METASCOPIA⁷. We further implemented model selection for the signals reaching epigenome-wide significance in the meta-analysis by using the ‘meta-Bayesian Information Criterion’ (meta-BIC) value. Following the notation from Bohning¹⁵, the likelihood is defined as

$$\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-1/2 (\hat{\lambda}_i - \lambda_i)^2 / \sigma_i^2\}$$

and the meta-likelihood as

$$\prod_{i=1}^k \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-1/2 (\hat{\lambda}_i - \lambda_i)^2 / \sigma_i^2\}$$

The value of the Bayesian Information Criterion is calculated as in Wit *et al.*¹⁶

$$BIC = \ln(n) k - 2 \ln(\hat{L})$$

where

\hat{L} = maximized value of likelihood function

n = (combined) sample size

k = number of parameters estimated by the model

Therefore, the value of the meta-log-likelihood can simply be computed as the sum of the log-likelihoods of the individual studies, and the ‘meta-BIC’ value for the meta-analysis result can subsequently be computed based on that sum, i.e., based on the meta-log-likelihood.

methylSCOPA and META-methylSCOPA

As methylSCOPA and META-methylSCOPA are the extensions for methylation data analysis of SCOPA and META-SCOPA⁷, respectively, installation procedures, command line options, and output columns for methylSCOPA and META-methylSCOPA are largely analogous to those for SCOPA and META-SCOPA, respectively and are detailed in the tutorial provided along the software.

Required file formats

Similarly to SCOPA⁷, methylSCOPA requires phenotype data in GEN/SAMPLE format utilized by the IMPUTE and SNPTESTv.2. The methylation data file format required by the software is described in the tutorial available online <http://www.imperial.ac.uk/people/h.draisma/research.html>.

Case study

Study populations

To implement our novel methylSCOPA and META-methylSCOPA methods and tools, we performed an MP-EWAS of two correlated glycaemic traits: FI and FG in two independent cohorts, followed by their meta-analysis. These two independent cohorts are the Northern Finland Birth Cohorts (NFBC) 1966/1986 which cover almost all births in the two northernmost provinces of Finland between the expected dates of delivery falling in 1966¹⁷ (N=12,058 live-born children) and between 1st of July 1985 and 30th June 1986¹⁸ (N=9,432 live-born children). The children born to the cohort have been followed up throughout their lives, and here, we used data from the 31-year clinical examination for NFBC1966 and from the 15/16 year clinical examination for the NFBC1986. FI and FG were measured after overnight fasting and processed according to the standard protocol. The methylation data were obtained from the Illumina Infinium

HumanMethylation450K BeadChip array for 807 randomly selected subjects that had provided fasting blood samples at the 31-year clinical examination (NFBC1966) and for 15-16-year-old individuals (NFBC1986). After quality control, the MP-EWAS included data for 643/457 individuals from NFBC1966 and NFBC1986, respectively. All the individuals included in the study have provided written informed consent (or parents of the participants of NFBC1986). The study was approved by the ethical committees of the University of Oulu and Imperial College London (Approval:18IC4421).

Quality control of the methylation data

Methylation data was quality controlled as follows: we 1) removed duplicate samples, 2) filtered based on methylation detection *P*-value, 3) performed subset quantile normalization of raw methylation signal intensity values¹⁹, 4) removed methylation data batch effects using the pipeline CPACOR²⁰, 5) detected and removed methylation data outliers as well as samples with gender mismatch, 6) applied white blood cell type composition correction, using “Houseman estimates”²¹, 7) transformed the values resulting from the previous step (step 6) to “beta” values, 8) applied inverse normal transformation to the methylation “beta” values, and finally 9) calculated residuals of the inverse normal-transformed methylation “beta” values by linear regression, using as covariates the Houseman estimates, first 30 control probe PC scores, participant sex, smoking status (current/ever/never as defined in keeping with the definitions as used for NFBC1966 and NFBC1986 within the CARTA consortium [Morris et al., BMJ Open 2015;5:e008808], and also including an additional category ‘unknown’ for participants with missing data on smoking status), and – in the case of Model 2 analysis – BMI, TG, and WHR.

Phenotype definitions, imputation and transformations

We used fasting plasma glucose values (mmol/l) and fasting circulating insulin values (pmol/l) as the main variables of interest in our MP-EWAS. FG was used as such whereas we applied natural log-transformation

to FI. We excluded from the analyses non-fasting individuals, pregnant women, diabetic individuals defined as fasting plasma glucose ≥ 7 mmol/l; 2-hour post oral glucose tolerance test glucose ≥ 11.1 mmol/L, HbA1c $\geq 6.5\%$; diagnosis of type 1 or type 2 diabetes; or on diabetes treatment [oral and insulin] and individuals on lipid-lowering medication (ATC code “A10”). We then imputed the phenotype data for males and females separately with random forest (MisForest in R). We included use of oral contraceptives for the imputation model for females and additionally for both we included other available indicators of metabolic health, such as blood pressure and metabolomic data. The data for both sexes were combined after imputation. Finally, we regressed out participant sex and smoking status (current/ever/never/unknown), and – in the case of Model 2, see below – BMI, TG, and WHR from the FI and FG values. In fixed-effects meta-analysis of NFBC1966 and NFBC1986 using the functions “escalc” and “rma” in the “R” package ‘metafor’, the Pearson’s product-moment correlation between the resulting FG and FI residuals used for Model 2 analysis was 0.23 ($P < 0.0001$).

MP-EWAS

We fitted two models: Model 1 allows for the detection of effects that might be shared across ‘phenotypes of interest’ (FG and FI in this case) and measured potential confounders, i.e. confounders were included in the same model. The confounders considered were triglycerides (TG), body mass index (BMI) and waist/hip ratio (WHR). We regressed out the effect of sex and smoking status (current/ever/never/unknown) from these confounders, similarly as we did for the measurements of FG and FI. Model 2 includes multi-phenotype EWAS of FI/FG corrected for measured potential confounders and allows for the detection of effects that are unique to the phenotypes of interest, i.e. which are not shared with the measured potential confounders. In other words, Model 2 had the measured potential confounders TG, BMI and WHR regressed out from the FG and FI values prior to model fitting. Mathematically, the two MP-EWAS models fitted were as follows:

1. Methylation probe “beta” residuals $value_i = \beta_0 + \beta_1 FGres_i + \beta_2 \ln(FI)res_i + \beta_3 BMIres + \beta_4 TGres + \beta_5 WHRres + e_i$
2. Methylation probe “beta” residuals $value_i = \beta_0 + \beta_1 FGres_i + \beta_2 \ln(FI)res_i + e_i$

where $i=1, \dots, n$, $FGres$ = fasting glucose residuals, $\ln(FI)res$ = residuals of natural log-transformed fasting insulin, $BMIres$ = body mass index, $TGres$ = fasting serum triglyceride level and $WHRres$ = waist/hip ratio.

We meta-analyzed the cohort-specific MP-EWAS results with META-methylSCOPA, mapped genomic locations to CGCh37/hg19, and adopted $P < 1 \times 10^{-7}$ to denote epigenome-wide significance. From the META-methylSCOPA meta-analysis results, we filtered out associations involving probes that could be cross-reactive, polymorphic, or have been suggested to be excluded from analysis for any other reason by Chen et al. [Chen et al., Epigenetics. 2013 Jan 11;8(2)] (their files “List of cross-reactive probes” and “List of polymorphic CpGs”), Naeem et al. [Naeem et al., 2017; Zhou et al., 2017] (column “MASK.general” as in their file “hm450.hg19.manifest.tsv.gz” and columns “MASK_general_FIN” and “MASK_general_EUR” as in their file “hm450.hg19.manifest.pop.tsv.gz” as on [“http://zwdzwd.github.io/InfiniumAnnotation#download”](http://zwdzwd.github.io/InfiniumAnnotation#download) [accessed 2019-05-19 BST]).

Results and discussion

Our MP-EWAS (**Figure 1**) of 466,342 methylation probes and FG and FI using methylSCOPA and META-methylSCOPA resulted in one epigenome-wide significant signal at cg13708645 (chr12:121,974,305; $P_{Model1} = 1.2 \times 10^{-8}$; $P_{Model2} = 0.48$) within *KDM2B* (**Table 1, Figure 1**), and one at suggestive ($P < 10^{-6}$) significance at cg05063096 (chr3:143,689,810; $P_{Model1} = 2.0 \times 10^{-7}$; $P_{Model2} = 2.3 \times 10^{-7}$) within *C3orf58* (**Table 1, Figure 2**).

The locus at *KDM2B* has the most pronounced association with FI, BMI and WHR in univariate analyses (**Table 2**). It was not significant in Model 2. Model selection for this locus though metaBIC identified BMI and WHR as the most significant phenotype combination (**Table 3**), the effect on FI at this locus was driven through its effect on obesity phenotypes. The locus at *C3orf58*, identified within Model 2 with suggestive significance, showed stronger effect (**Table 1**), when additional phenotypes were considered in Model 1. The strongest phenotype effects at *C3orf58* were on combination of FI and WHR, while the effect on FI in univariate analysis would have not led to the detection of this epigenetic effect.

The univariate analyses showed that these signals would not have been detected in traditional EWAS for each trait separately, thus, indicating the improved power from the joint analysis of the correlated traits. We also characterized several established FG/FI or other relevant phenotype-associated EWAS signals through our MP-EWAS approach. A recent large-scale study in 4,808 (discovery) and 11,750 (replication) non-diabetic individuals by Liu *et al.* reported nine novel differentially methylated sites in whole blood with $P < 1.27 \times 10^{-7}$: sites in *LETM1*, *RBM20*, *IRS2*, *MAN2A2* genes and 1q25.3 region were associated with FI; sites in *FCRL6*, *SLAMF1*, *APOBEC3H* genes and 15q26.1 region were associated with FG²². We show that of the sites associated with FI, 1q25.3 and *FAM92B* (not replicated in the original study) were also nominally ($P < 0.05$) associated in our study, (**Table 4**). Similarly of the FG-associated sites, *BRE* (not replicated in the original study) was nominally associated in our study (**Table 4**).

Another study by Zaghlool *et al.*²³ aimed at elucidating the molecular pathways of 20 previously established CpG sites by using multi-omics data in 359 samples from the multi-ethnic Qatar Metabolomics Study on Diabetes. We observe associations at nominal significance at six of these 20 sites and demonstrate associations within Model 1 at *PHGDH*, *TXNIP*, *SLC7A11*, *CPT1A*, *MYO5C* and *ABCG1* through the dissection of multi-phenotype effects within our relatively small study sample (**Table 5**).

Conclusions

We have extended the multi-phenotype analyses to the EWAS framework and implemented this in the publicly available software tools methylSCOPA and METAmethylSCOPA. The application of the method to glycaemic traits demonstrated its enhanced power over single-trait EWAS for correlated phenotypes in large-scale data and the ability to characterize signals that are associated with correlated phenotypes.

Availability and requirements

Project name: methylSCOPA.

Availability: the methylSCOPA and META-methylSCOPA tutorial can be found on

“<http://www.imperial.ac.uk/people/h.draisma/research.html>”. The methylSCOPA and META-

methylSCOPA software can be found on “<http://doi.org/10.5281/zenodo.1137744>” and on

“<https://doi.org/10.5281/zenodo.1286392>”, respectively.

Operating system(s): Linux.

Programming language: C++ (including files from the ALGLIB project for statistical analysis and the TCLAP project for command line argument parsing).

Any restrictions on use by academics: none.

Acknowledgments

Northern Finland Birth Cohort (NFBC1966) would like to thank the late professor Paula Rantakallio (launch of NFBC1966), the participants in the 31 year study and the NFBC project center.

Funding

This project was funded by the Wellcome Trust (WT205915) seed award in science to IP. NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 2/97, 8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 160/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231.

This work used the computing resources of the UK MEDical BIOinformatics partnership - aggregation, integration, visualisation and analysis of large, complex data (UK MED-BIO) which is supported by the Medical Research Council [grant number MR/L01632X/1]; the Imperial College High Performance Computing Service, URL: <http://www.imperial.ac.uk/admin-services/ict/self-service/research-support/hpc/>.

Availability of data and materials

The methylSCOPA and METAmethylSCOPA software tools are freely available at URL: xxx.

The Northern Finland Birth Cohort data which were used for the application of the developed tool are available upon collaboration and formal data request only, please see <http://www.oulu.fi/nfbc/node/18136>.

Competing interests

None.

References

1. Amos, C. I. & Laing, A. A comparison of univariate and multivariate tests for genetic linkage. *Genet. Epidemiol.* **10**, 671–6 (1993).
2. Allison, D. B. *et al.* Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *Am. J. Hum. Genet.* **63**, 1190–1201 (1998).
3. Banerjee, S., Yandell, B. S. & Yi, N. J. Bayesian quantitative trait loci mapping for multiple traits. *Genetics* **179**, 2275–2289 (2008).
4. Kim, S. & Xing, E. P. Statistical Estimation of Correlated Genome Associations to a Quantitative Trait Network. *PLoS Genet.* **5**, e1000587 (2009).
5. Jiang, C. & Zeng, Z. B. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**, 1111–27 (1995).
6. Shriner, D. Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies. *Front. Genet.* **3**, 1 (2012).
7. Mägi, R. *et al.* SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide association studies of multiple correlated phenotypes. *BMC Bioinformatics* **18**, 25 (2017).
8. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
9. Fraga, M. F. *et al.* Epigenetic differences arise during the lifetime of monozygotic twins. **102**, 10–15 (2005).
10. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).

11. Kriebel, J. *et al.* Association between DNA Methylation in Whole Blood and Measures of Glucose Metabolism: KORA F4 Study. *PLoS One* **11**, (2016).
12. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet. Diabetes Endocrinol.* **3**, 526–534 (2015).
13. Lin, X., Barton, S. & Holbrook, J. D. How to make DNA methylome wide association studies more powerful. **8**, 1117–1129 (2016).
14. Becker, B. J. & Wu, M. The Synthesis of Regression Slopes in Meta-Analysis. *Stat. Sci.* **22**, 414–429 (2007).
15. Böhning, D. Meta-Analysis. A Unifying Meta-Likelihood Approach Framing Unobserved Heterogeneity, Study Covariates, Publication Bias, and Study Quality. *Methods Inf Med* **44**, 127–135 (2005).
16. Wit, E., van den Heuvel, E. & Romeijn, J.-W. ‘All models are wrong ...’: an introduction to model uncertainty. *Stat. Neederlandica* **66**, 217–236 (2012).
17. Rantakallio P. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand.* **193**, 1+ (1969).
18. Järvelin, M. R. *et al.* Ecological and individual predictors of birthweight in a northern Finland birth cohort 1986. *Paediatr. Perinat. Epidemiol.* **11**, 298–312 (1997).
19. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325–341 (2012).

20. Lehne, B. *et al.* A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.* **16**, 1–12 (2015).
21. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, (2012).
22. Liu, J. *et al.* Novel DNA methylation sites of glucose and insulin homeostasis: an integrative cross-omics analysis. *bioRxiv* 1–40 (2018).
23. Zaghlool, S. B. *et al.* Deep molecular phenotypes link complex disorders and physiological insult to CpG methylation. *Hum. Mol. Genet.* **27**, 1106–1121 (2018).

Tables

Table 1. Lead cgIDs in META-methylSCOPA meta-analysis of fasting glucose (FG) and fasting insulin (FI) in 1,100 individuals from NFBC1966 and NFBC1986 in Models 1 and 2. Measured potential confounders body mass index (BMI), triglycerides (TG) and waist-to-hip ratio (WHR) were included in Model 1 whereas their effects were regressed out in Model 2.

Model	Locus	Lead cgID	Chr	Position ^a (bp)	FG effect (SE)	FI effect (SE)	BMI effect (SE)	TG effect (SE)	WHR effect (SE)	P-value
1	<i>KDM2B</i>	cg13708645	12	121,974,305	-0.002 (0.003)	0.004 (0.004)	0.001 (4.6×10⁻⁴)	-4.7×10⁻⁴ (0.002)	0.075 (0.027)	1.2×10⁻⁸
1	<i>C3orf58</i>	cg05063096	3	143,689,810	0.002 (7.9×10 ⁻⁴)	-0.006 (0.001)	7.0×10 ⁻⁵ (1.2×10 ⁻⁴)	0.002 (6.4×10 ⁻⁴)	0.022 (0.007)	2.0×10 ⁻⁷
2	<i>KDM2B</i>	cg13708645	12	121,974,305	-0.002 (0.003)	0.005 (0.004)				0.48
2	<i>C3orf58</i>	cg05063096	3	143,689,810	0.002 (7.9×10⁻⁴)	-0.006 (0.001)				2.3×10⁻⁷

Chr: chromosome. SE: standard error. FG: fasting plasma glucose level. FI: fasting circulating insulin level. BMI: body mass index. TG: fasting serum triglyceride level. WHR: waist-to-hip ratio

^aPosition reported for NCBI build GRCh37 (UCSC hg19 assembly)

Table 2. Univariate EWAS meta-analysis at lead cgIDs of FI and FG in 1,100 individuals from NFBC1966 and NFBC1986. Measured potential confounders body mass index (BMI), triglycerides (TG) and waist-to-hip ratio (WHR) were included in Model 1 whereas their effects were regressed out in Model 2.

Model	Locus	Lead cgID	Chr	Position ^a (bp)	FG		FI		BMI		TG		WHR	
					Effect (SE)	P-value	Effect (SE)	P-value	Effect (SE)	P-value	Effect (SE)	P-value	Effect (SE)	P-value
1	<i>KDM2B</i>	cg13708645	12	121,974,305	0.002 (0.003)	0.60	0.013 (0.004)	1.7×10⁻⁴	0.002 (4×10⁻⁴)	2.6×10⁻⁹	0.005 (0.002)	0.044	0.128 (0.023)	1.9×10⁻⁸
1	<i>C3orf58</i>	cg05063096	3	143,689,810	0.001 (8×10 ⁻⁴)	0.11	-0.003 (9×10 ⁻⁴)	0.002	1×10 ⁻⁴ (1×10 ⁻⁴)	0.39	9×10 ⁻⁴ (6×10 ⁻⁴)	0.14	0.016 (0.006)	0.006
2	<i>KDM2B</i>	cg13708645	12	121,974,305	-0.001 (0.003)	0.71	0.004 (0.004)	0.30						
2	<i>C3orf58</i>	cg05063096	3	143,689,810	0.001 (8×10⁻⁴)	0.16	-0.005 (0.001)	1.0×10⁻⁶						

Chr: chromosome. SE: standard error. FG: fasting plasma glucose level. FI: fasting circulating insulin level. BMI: body mass index. TG: fasting serum triglyceride level. WHR: waist-to-hip ratio

^aPosition reported for NCBI build GRCh37 (UCSC hg19 assembly)

Table 3. Dissection of the multiple phenotype association signals for the lead cgIDs in the MP-EWAS of fasting glucose (FG) and fasting insulin (FI) in 1,100 individuals from NFBC1966 and NFBC1986 in

Models 1 and 2. The values displayed are the differences in metaBIC from that of the null model.

Potential confounders of body mass index (BMI), triglycerides (TG) and waist-to-hip ratio (WHR) were included in Model 1 whereas their effects were regressed out in Model 2.

Model	Model 1 cg13708645	Model 1 cg05063096	Model 2 cg13708645	Model 2 cg05063096
BMI	-28.26	6.10		
BMI+FG	-26.33	10.59		
BMI+FG+FI	-21.73	-2.57		
BMI+FG+TG	-21.23	15.91		
BMI+FG+TG+FI	-16.57	-2.63		
BMI+FG+TG+WHR	-23.45	14.77		
BMI+FG+TG+WHR+FI	-17.88	-8.40		
BMI+FG+WHR	-28.67	8.95		
BMI+FG+WHR+FI	-23.11	-8.85		
BMI+FI	-23.87	-4.17		
BMI+TG	-23.21	11.51		
BMI+TG+FI	-18.35	-3.33		
BMI+TG+WHR	-25.69	10.29		
BMI+TG+WHR+FI	-20.15	-8.57		
BMI+WHR	-30.83	4.43		
BMI+WHR+FI	-25.69	-9.87		
FG	0.18	3.98	1.70	4.68
FG+FI	-8.47	-3.53	7.09	-18.33
FG+TG	2.47	8.97		
FG+TG+FI	-2.85	-5.31		
FG+TG+WHR	-19.33	9.73		
FG+TG+WHR+FI	-16.97	-15.23		
FG+WHR	-24.01	3.65		
FG+WHR+FI	-22.47	-15.55		
FI	-9.72	-4.96	5.66	-18.90
TG	1.98	4.78		
TG+FI	-3.75	-5.81		
TG+WHR	-20.71	5.07		
TG+WHR+FI	-18.89	-15.27		
WHR	-25.22	-1.00		

WHR+FI	-24.71	-16.41		
--------	--------	--------	--	--

BMI, body mass index; TG, triglycerides; WHR, waist-to-hip ratio.

Table 4. Sites reaching epigenome-wide significance in the study by Liu *et al.*²² and their associations

in META-methylSCOPA meta-analysis within our study sample including 1,100 individuals from

NFBC1966 and NFBC1986. Measured potential confounders body mass index (BMI), triglycerides (TG)

and waist-to-hip ratio (WHR) were included in Model 1 whereas their effects were regressed out in

Model 2.

Locus	CpG	Chr	Position	Regulatory feature	Trait	Model 1 P-value	Model 2 P-value
<i>FCRL6</i> #	cg00936728	1	159772194	NA	Glucose	0.45	0.24
<i>SLAMF1</i> #	cg18881723	1	160616870	Promoter associated	Glucose	0.74	0.68
<i>1q25.3</i> #	cg13222915	1	184598594	NA	Insulin	0.014	0.24
<i>BRE</i>	cg20657709	2	28509570	NA	Glucose	0.091	0.039
<i>LRPPRC</i>	cg01913188	2	44223249	Promoter associated	Glucose	0.38	0.87
<i>IRAK2</i>	cg14527942	3	10276383	NA	Insulin	0.31	0.69
<i>LETM1</i> #	cg13729116	4	1859262	Promoter associated	Insulin	0.41	0.53
<i>RBM20</i> #	cg15880704	10	112546110	NA	Insulin	0.52	0.61
<i>IRS2</i> #	cg25924746	13	110432935	Promoter associated	Insulin	0.014	0.77
<i>SPTB</i>	cg07119168	14	65225253	NA	Glucose	0.84	0.33
<i>15q26.1</i> #	cg18247172	15	91370233	NA	Glucose	0.025	0.05
<i>MAN2A2</i> #	cg20507228	15	91460071	Promoter associated (Cell type specific)	Insulin	0.13	0.10
<i>FAM92B</i>	cg06709610	16	85143924	NA	Insulin	0.077	0.080
<i>CD300A</i>	cg08087047	17	72461209	NA	Glucose	0.51	0.28
<i>APOBEC3</i> #	cg06229674	22	39492189	NA	Glucose	0.63	0.77

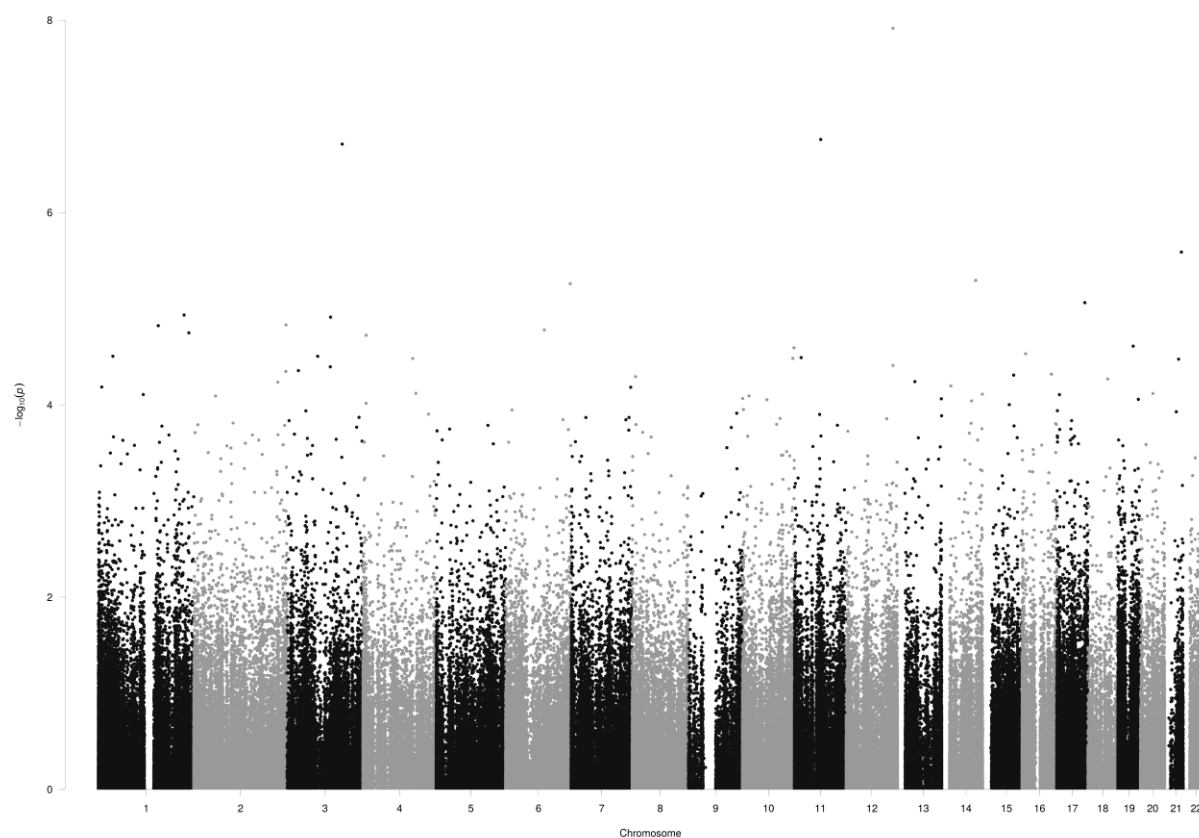
#Replicated in Liu *et al.*

Table 5. Previously established sites for diabetes and its risk factors and their associations within our study sample including 1,100 individuals from NFBC1966 and NFBC1986. Measured potential confounders body mass index (BMI), triglycerides (TG) and waist-to-hip ratio (WHR) were included in Model 1 whereas their effects were regressed out in Model 2.

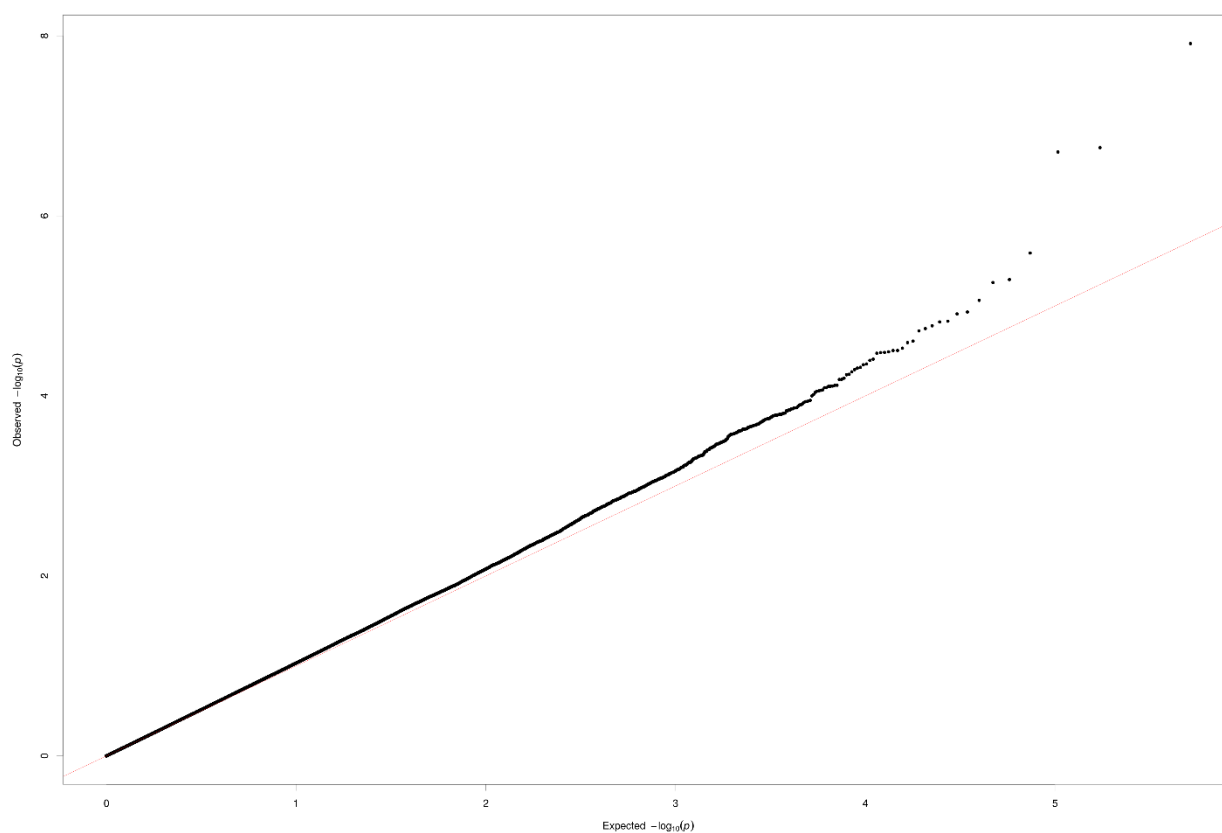
Locus	Trait	CpG	Chr	Position	Model 1 P-value	Model 2 P-value
<i>DHCR24</i>	Obesity	cg17901584	1	55353706	0.075	0.53
<i>GFI1</i>	Smoking	cg09935388	1	92947588		
<i>PHGDH</i>	BMI, blood pressure, liver function	cg14476101	1	120255992	0.0056	0.031
<i>TXNIP</i>	T2D	cg19693031	1	145441552	8.36E-04	0.051
	Metabolite, lipid	cg23079012	2	8343710	0.94	0.55
<i>ALPPL2</i>	Smoking	cg21566642	2	233284661	0.25	0.18
<i>UGT2B15</i>	Metabolite	cg09189601	4	69514031	0.56	0.19
<i>SLC7A11</i>	Obesity	cg06690548	4	139162808	0.0037	0.44
<i>AHRR</i>	Smoking	cg05575921	5	373378	0.40	0.12
	Metabolite, protein, smoking	cg06126421	6	30720080	0.26	0.45
<i>LOC1001323546</i>	BMI, blood pressure	cg18120259	6	43894639	0.15	0.087
<i>SLC25A22</i>	Metabolite, T2D, BMI	cg09441501	11	798350	0.56	0.67
<i>CPT1A</i>	T2D and obesity	cg00574958	11	68607622	1.74E-07	0.91
		cg13526915	14	24164078		
<i>MYO5C</i>	Obesity	cg06192883	15	52554171	6.53E-04	0.48
<i>TPM1</i>	Metabolite, smoking	cg10403394	15	63349192	0.32	0.85
<i>RARA</i>	Smoking	cg19572487	17	38476024	0.72	0.93
<i>F2RL3</i>	Smoking	cg03636183	19	17000585	0.42	0.56
<i>SLC1A5</i>	BMI, blood pressure	cg22304262	19	47287778	0.33	0.87
<i>ABCG1</i>	T2D and obesity	cg06500161	21	43656587	2.58E-06	0.077

Figures

A)



B)



c)

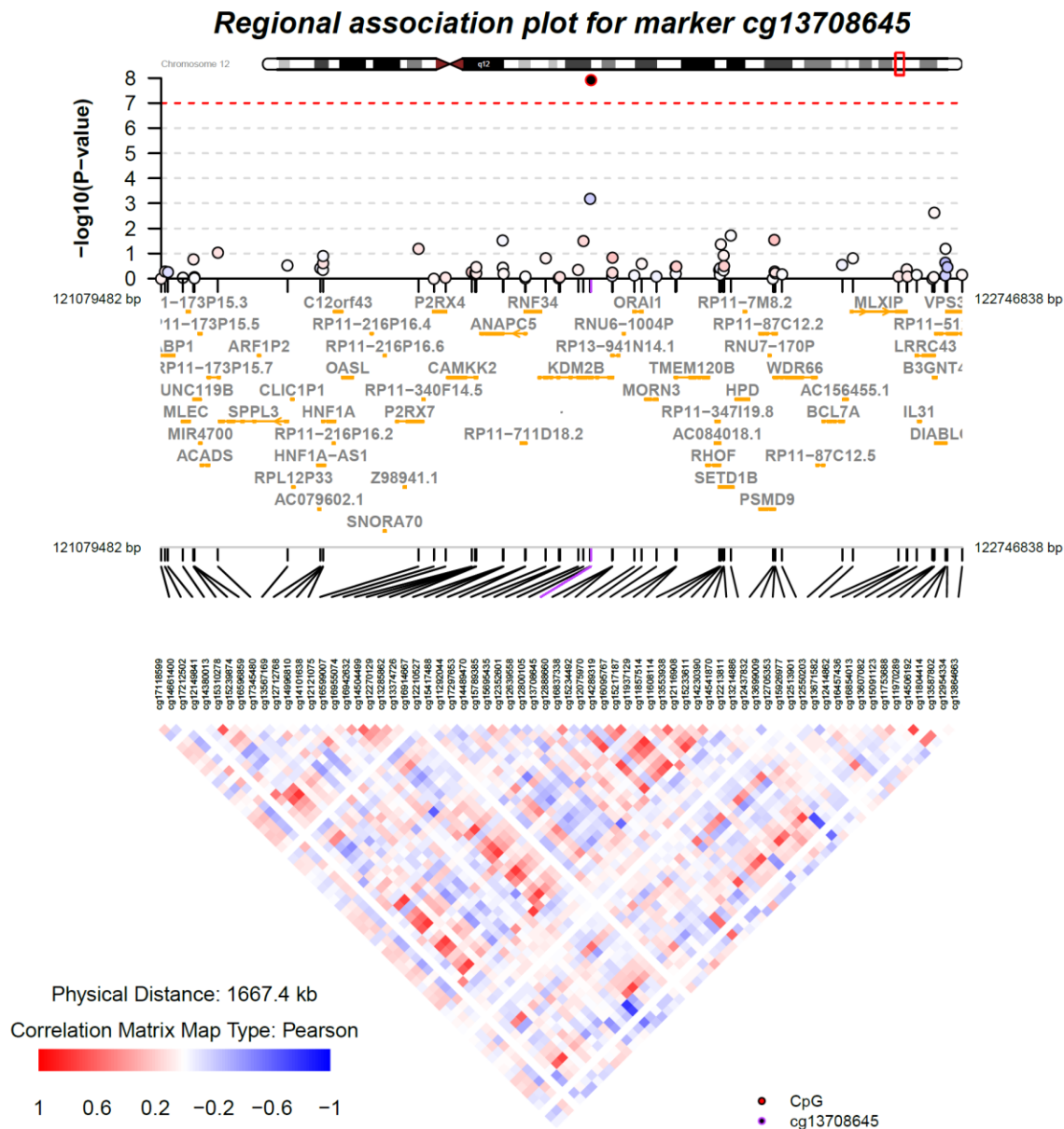
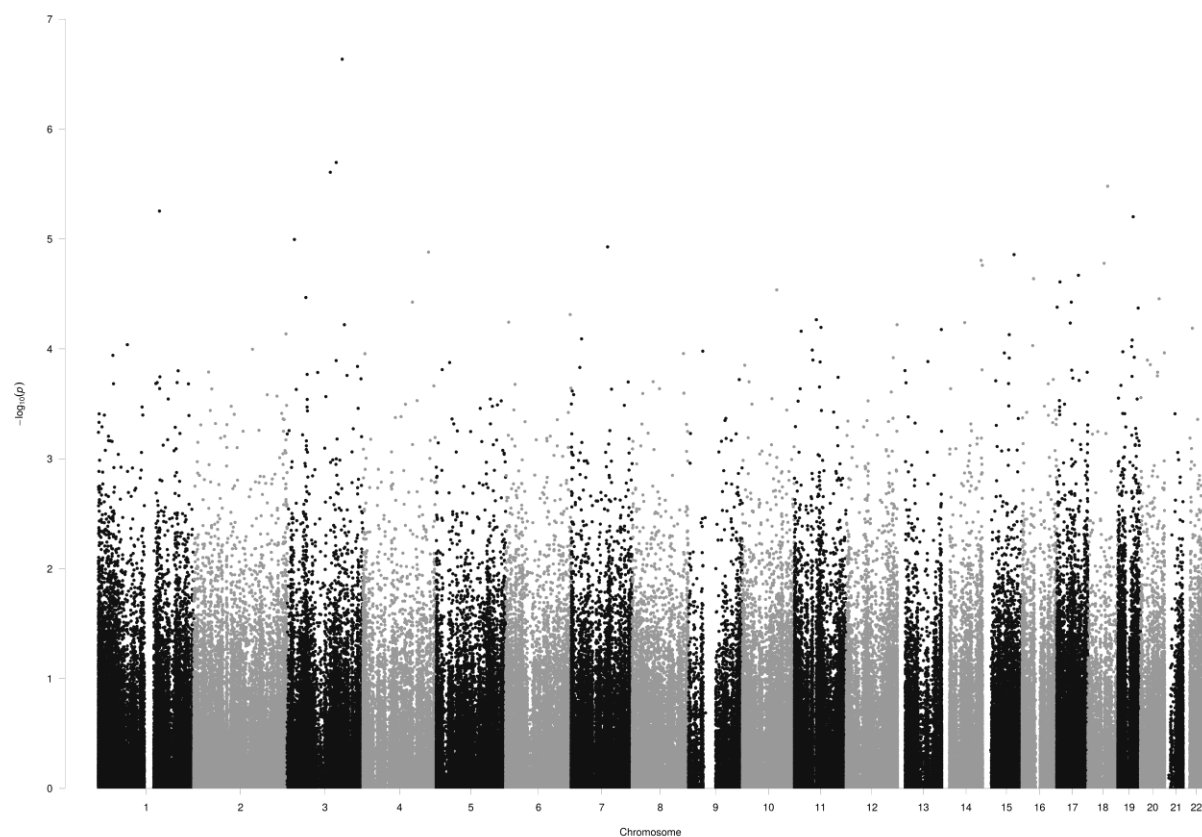


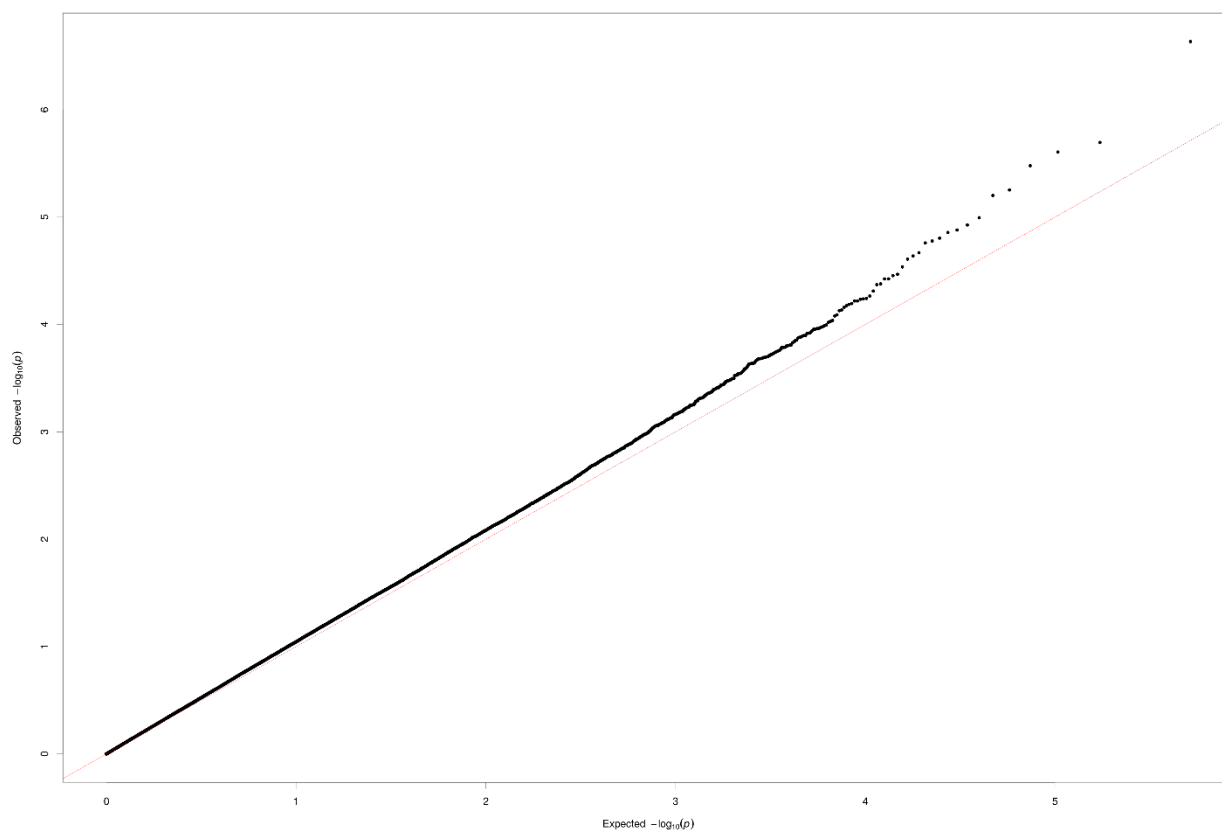
Figure 1. Results from the meta-analysis of MP-EWAS for FG and FI from Model 1. A) Manhattan plot, B) QQ-plot, and C) Regional association plot for the top signal. In the regional association plot, each point represents a CpG passing quality control in the association analysis, plotted with their p -value (on a $-\log_{10}$ scale) as a function of genomic position (NCBI build GRCh37, UCSC hg 19 assembly). The lead CpG is represented by the circle with the red edge and black face. The color coding of all other CpGs

indicates Pearson correlation with the lead CpG in meta-analysis of NFBC1966 age 31 and NFBC1986 data. We performed the meta-analysis of the correlations between CpGs using function “escalc” from the ‘metafor’ package in the statistical language and environment “R” [R package], and created the signal plots using the coMET package. Gene annotations are taken from the University of California Santa Cruz genome browser

A)



B)



CpG is represented by the circle with the red edge and black face. The color coding of all other CpGs indicates Pearson correlation with the lead CpG in meta-analysis of NFBC1966 age 31 and NFBC1986 data. We performed the meta-analysis of the correlations between CpGs using function “escalc” from the ‘metafor’ package in the statistical language and environment “R” [R package], and created the signal plots using the coMET package. Gene annotations are taken from the University of California Santa Cruz genome browser