Supplementary Note 1

ALS - King's College London (Batches 1 & 2)

These data are part of ProjectMinE[1]. In brief, project MinE is a collaboration of (inter)national groups with the aim to collect 22,500 DNA profiles to investigate rare and common (epi)genetic variation contributing to the development of Amyotrophic Lateral Sclerosis (ALS). The participants of this study consisted of a subset of 1433 individuals of UK nationality from the UK National DNA Bank for MND Research who were put forward for DNA methylation (DNAm) profiling. Cases were diagnosed with ALS in one of 20 UK hospitals by neurologists specialized in motor neuron diseases; patients had no family history for ALS and were of self-reported European descent. All cases and controls gave written informed consent. The institutional review board at King's College London approved this study.

DNAm assays

DNA was extracted by use of standard methods at three centres within 1 week of the blood being drawn (usually on the same day) and was stored centrally at the UK DNA banking network in Manchester. We used a barcode-based sample tracking system to minimise the risk of clerical error. DNAm status of the participants was extracted from whole blood samples using Illumina Infinium HumanMethylation450 BeadChip array following manufacturer's protocol. As these samples were run in two separate batches at two different time points (batch 1 n = 666; batch 2 n = 767), they have been included as two separate cohorts. Both cohorts followed the same quality control (QC) pipeline detailed as follows. Idat files for all samples were imported to the R environment using the methlumIDAT() function from the *methylumi* package[2]. These underwent a stringent QC pipeline. First, the level of the M(ethylated) and U(nmethylated) signal intensities were examined and samples with a median < 1500 were excluded. Second, the DNAm beta values were uploaded to the Epigenetic Clock software[3] to predict tissue type; all samples were predicted of blood origin. Third, using the ten control probes included on the 450K array, a measure of the efficiency of the sodium bisulfite conversion reaction was calculated; all samples had a "conversion score" > 90. Fourth, the 65 SNP probes included on the 450K array were correlated between all pairs of samples to identify genetically identical samples or duplicates; any duplicates were excluded (n = 4). Fifth, multidimensional scaling was performed for DNAm probes on each of the sex chromosomes and compared to the reported gender; 23 samples were excluded for wither clustering with a different sex to that reported in the phenotype file or clustering to females on the X chromosome and males on the Y chromosome. Sixth, 1348 samples (94.1%) had genotype data available for 34/65 SNPs present on the 450K array, comparing these variants across the two platforms; all samples were concordant. The data were then processed with the pfilter() function from the *wateRmelon* package[4] excluding 1 sample with >1% of sites with a detection p > 0.05.

Genotyping assays

Genotyping was performed as previously described[5].

Data availability

Data are available to researchers by request as outlined in the Project MinE access policy.

Acknowledgements

The authors want to thank the study participants that contributed whole blood for this study and the Project MinE GWAS Consortium.

Funding

This is in part an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND - www.jpnd.eu *(United Kingdom, Medical Research Council* (MR/L501529/1; MR/R024804/1) *and Economic and Social Research Council* (ES/L008238/1)*)* and through the Motor Neurone Disease Association. This study represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. Samples used in this research were in part obtained from the UK National DNA Bank for MND Research, funded by the MND Association and the Wellcome Trust. We acknowledge sample management undertaken by Biobanking Solutions funded by the Medical Research Council at the Centre for Integrated Genomic Medical Research, University of Manchester.

Accessible Resource for Integrative Epigenomic Studies (ARIES)

Samples were drawn from the Avon Longitudinal Study of Parents and Children[6,7]. Pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were invited to take part in the study. The total number of pregnancies enrolled is 15,247 pregnancies, resulting in 15,458 foetuses. Of this total sample of 15,458 foetuses, 14,775 were live births and 14,701 were alive at 1 year of age. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool" and reference the following webpage: http://www.bristol.ac.uk/alspac/researchers/our-data/.

Blood from 1018 mother–child pairs (children at three time points and their mothers at two time points) were selected for analysis as part of the Accessible Resource for Integrative Epigenomic Studies (ARIES, http://www.ariesepigenomics.org.uk/)[8]. Following DNA extraction, samples were bisulphite converted using the Zymo EZ DNA Methylation™ kit (Zymo, Irvine, CA, USA).

Written informed consent has been obtained for all ALSPAC participants. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

DNAm assays

Following conversion, genome-wide DNAm was measured using the Illumina Infinium HumanMethylation450 (HM450) BeadChip. The arrays were scanned using an Illumina iScan, with initial quality review using GenomeStudio. ARIES was preprocessed and normalised using the *meffil* R package.[9] ARIES consists of 5469 DNAm profiles obtained from 1022 mother-child pairs measured at five time points (three time points for children: birth, childhood and adolescence; and two for mothers: during pregnancy and at middle age). Low quality profiles were removed from further processing, and the remaining 4593 profiles were normalised using the Functional Normalization algorithm[10] with the top 10 control probe principal components (PCs). Full details of the preprocessing and normalization of ARIES has been described previously[9].

Genotyping assays

The ARIES participants were previously genotyped as part of the larger ALSPAC study, with QC, cleaning and imputation performed at the cohort level before extraction of the subset comprising ARIES. Children were genotyped using the Illumina HumanHap550 quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA) by the Wellcome Trust Sanger Institute (WTSI; Cambridge, UK) and the Laboratory Corporation of America (LCA, Burlington, NC, USA). Individuals were excluded on the basis of incorrect gender assignment, abnormal heterozygosity (<0.320 or >0.345 for WTSI data; <0.310 or >0.330 for LCA data), high missingness (>3 %), cryptic relatedness (>10 % identity by descent) and non-European ancestry (detected by multidimensional scaling analysis). Following QC, the final directly genotyped dataset contained 500,527 SNP loci. Mothers were genotyped using the Illumina Human660W-quad genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA) at the Centre National de Génotypage (CNG; Paris, France). Individuals were excluded based on non-European ancestry, missingness, relatedness, gender mismatches and heterozygosity. PLINK (v1.07)[11] was used to carry out QC measures on an initial set of 10,015 subjects (including non-ARIES ALSPAC participants) and 557,124 directly genotyped SNPs. Following QC, the final directly genotyped dataset contained 526,688 SNP loci. Imputation was performed to increase the SNP density for all genotyped mothers and children combined. Genotypes were phased together using SHAPEIT (version 2, revision 727)[12] and then imputed against the 1000 Genomes reference panel (phase 1, version 3, phased using SHAPEIT version 2, December 2013, using all populations[13]) using IMPUTE (v2.2.2)[14]. Genotypes were filtered to have Hardy Weinberg equilibrium (HWE) $p > 5e{-}7$, MAF >1 % and imputation info score >0.8.

Data availability

Data are available to researchers by request from the Avon Longitudinal Study of Parents and Children Executive Committee (http://www.bristol.ac.uk/alspac/researchers/access/) as outlined in the study's access policy http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf.

Acknowledgements

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. 450K DNAm array data and part of the genotype data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Funding

This work was supported by the UK Medical Research Council; Wellcome (www.wellcome.ac.uk; [grant number 102215/2/13/2 to ALSPAC]); the University of Bristol to ALSPAC; the UK Economic and Social Research Council (www.esrc.ac.uk; [ES/N000498/1] to CR) and the UK Medical Research Council (www.mrc.ac.uk; grant numbers [MC_UU_00011/1, MC_UU_00011/5 to JLM, GH, GDS, CLR and MS]). DNAm data in the ALSPAC cohort were generated as part of the UK BBSRC funded (BB/I025751/1 and BB/I025263/1) Accessible Resource for Integrated Epigenomic Studies (ARIES, http://www.ariesepigenomics.org.uk) and was funded by MC_UU_00011/5 to CLR. GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. GWAS data in the mothers was funded by Wellcome WT088806. This publication is the work of the authors and Josine Min will serve as guarantor for the contents of this paper.

The BAMSE Study

The BAMSE (Swedish abbreviation for Children, Allergy, Milieu, Stockholm, Epidemiology) study is a Swedish longitudinal prospective birth cohort study of 4,089 children born between 1994 and 1996 in the area of Stockholm[15]. At ages 1, 2, 4, 8, 12 and 16 years, parents completed questionnaires on their children's health including allergic symptoms and diseases. At ages 4,8 and 16 years, a clinical follow-up (including blood sampling for DNA extraction) was performed. Written study consent was obtained from all participating children and their parents, and the study was approved by the Regional Ethics Committee in Stockholm (dnr 02-420 and 2010/1474/-31/3).

DNAm assays

Epigenome-wide DNAm was measured in 472 Caucasian children in BAMSE, using DNA extracted from blood samples collected at the age of 8 years (mean age 8.32 years) as previously described[16]. An aliquot (500 ng) of DNA per sample underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, USA). Samples were processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA). QC of analysed samples was performed using standardized criteria[16]. Samples were excluded (n=8) in case of sample call rate <99%, color balance >3, low staining efficiency, poor extension efficiency, poor hybridization performance, low stripping efficiency after extension and poor bisulfite conversion. After QC, 460 samples remained with both SNP and DNAm data available for analyses. Probes with a single nucleotide polymorphism in the single base extension site with a frequency of >5% were

excluded, as were probes with non-optimal binding, and the probe belonging to chromosome X and chromosome Y, resulting in the exclusion of 46,799 probes, leaving a total of 438,713 probes in the analysis. Furthermore, we implemented “DASEN” recommended from *wateRmelon* package[4] to do signal correction and normalization.

Genotyping assays

For genotyping, the Illumina Human 610-quad array (Illumina, Inc., San Diego, CA) was used. A total of 505 samples were genotyped (a subset of the study consisting of asthma cases and controls[17], out of which 485 samples were of good genotype quality. For imputation, the samples were excluded if their genotyping success rate was lower than 95% and SNPs were excluded for - Call rate < 95%; HWE p< 1e-6; MAF < 0.01. These QCed SNPs were imputed using Minimac (release stamp 2012-11-16)[18] based upon the GIANT ALL reference panel, phase 1 v3.20101123.

Data availability

Data from the BAMSE study is available from the GABRIEL consortium[17] as well as from our study portal at http://ki.se/en/imm/medallomics.

Funding

This work was supported by the Swedish Research Council (VR), Stockholm County Council (ALF), the Swedish Heart and Lung Foundation, the European Commission’s Seventh Framework Program MeDALL under grant agreement No 261357 and GABRIEL (No 018996), The Swedish Research Council Formas and Swedish Foundation for Strategic Research (the Epigene study). Erik Melén is supported by a grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement nº 757919, TRIBAL).

Acknowledgements

We would sincerely like to thank all participating children and their parents for their contribution in the BAMSE study.

Database of ischemic stroke from Hospital del Mar (BASICMAR)

BASICMAR is a prospective registry of stroke patients recruited in Hospital del Mar in Barcelona (Spain)[19]. The BASICMAR Register prospectively recruited all consenting patients who were admitted to our hospital from 2005 to 2017 (n=7200) with a diagnosis of stroke fulfilling World Health Organization criteria. From those enrolled in BASICMAR register, 598 blood DNA samples of ischemic stroke patients were selected for analysis. Inclusion criteria in BASICMAR cohorts were as follows: first-ever IS, brain imaging with CT or MRI and availability of the clinical data supporting the assigned stroke subtype according to TOAST classification[20]. All patients were assessed and classified by a neurologist and were included in the study by consecutive order of recruitment. All subjects were of European descent. The study was approved by the local ethics committee, CEIC-Parc de Salut Mar, and participants

gave written informed consent. The study was conducted according to the principles expressed in the Declaration of Helsinki and relevant legislation in Spain.

DNAm assays

Genomic DNA was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Orange, CA, USA) according to the manufacturer's procedure, with the alternative incubation conditions recommended when using the Illumina DNAm Assay. Genome-wide DNAm was assessed using the Illumina HumanMethylation450 Beadchip (Illumina Netherlands, Eindhoven, Netherlands) following the manufacturer's protocol with no modifications in Progenika Biopharma in Bizkaia, Spain. The arrays were scanned with the Illumina HiScan SQ scanner.

Initial QC of sample data is conducted using GenomeStudio version 2011.1 (Illumina, San Diego, CA, USA) with the DNAm module (version 1.9.0) to determine the status of staining, extension, hybridization, target removal, bisulfite conversion, specificity, non-polymorphic and negative controls (without background correction or normalization) as described previously[21]. We use all the samples that have a detection rate over 95%.

Sample QC and normalization was completed using with *meffil*[9] in R version 3.2.0. Samples that perform poorly in these QC were excluded from further analysis. We exclude all probes that were represented by a bead count <3 in at least 5% of the samples. DNAm sites having 1% of samples with a detection p of >0.05 were removed and cross-reactive probes were excluded, as well. Before analysis, DNAm values were corrected for background values and then normalized using functional normalization[9,10,21]. Finally, 529 samples passed QC.

Genotyping assays

The samples were genotyped using Illumina HumanOmni5PlusExome BeadChip and HumanCoreExome-12v1-1. PLINK (v1.07)[11] was used to carry out QC measures of the directly genotyped SNPs. Following QC, the final directly genotyped dataset contained 311.994 SNPs. Individual samples were tested for missing (<0.05), heterozygosity (>-0.3) and genetic sex discrepancies. We also tested for individual IBS, performing MDS plots with PCs. PCs were used to identify and remove ethnic outliers, and to adjust for population stratification in the downstream analyses. SNPs with genotype call rate <95% and/or a HWE with p<1e−6 were excluded, together with individuals with >5% missing genotypes. Samples that perform poorly in these QCs are excluded from further analysis.

Imputation was performed using IMPUTE v2.3.0[14] with the 1000 Genomes reference panel (phase 1, version 3, CEU SNP panel reference)[13]. We removed SNPs with imputation certainty <85% and r^2<30%, and also those SNPs with MAF <0.01. The final imputed dataset contained 9.081.881 SNPs.

Data availability

DNAm data is available in the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE69138.

Acknowledgements

We thank the whole neurovascular group from Hospital del Mar-IMIM (Barcelona, Spain): Marina Mola-Caminal, Ángel Ois, Ana Rodríguez-Campello, Carla Avellaneda, Rosa M Vivanco-Hidalgo, Elisa Cuadrado-Godia and Gemma Romeral.

Funding

This work was supported by the Agència de Gestió Ajuts Universitaris de Recerca (2014 SGR 1213); Spain's Ministry of Health (Ministerio de Sanidad y Consumo) through the Carlos III Health Institute (ISCIII-FIS-FEDER-ERDF, PI051737; PI12/01238, PI15/00451, PI15/00445); INVICTUS-PLUS, Instituto de Salud Carlos III RETIC (RD16/0019/0002); RecerCaixa 2013 research grant (JJ086116) and Fundació la Marató TV3 (76/C/2011).

Born in Bradford (BiB)

Born in Bradford is a longitudinal multi-ethnic birth cohort study aiming to examine the impact of environmental, psychological and genetic factors on maternal and child health and wellbeing[22]. Bradford is a city in the North of England with high levels of socio-economic deprivation and ethnic diversity. Women were recruited at the Bradford Royal Infirmary at 24-29 weeks gestation. For those consenting, a baseline questionnaire was completed. The full BiB cohort recruited 12,450 women comprising 13,773 pregnancies and 13,858 children between 2007 and 2010. Results of an oral glucose tolerance test (OGTT) and lipid profiles were obtained on the mothers during pregnancy at recruitment (24-29 weeks gestation), and pregnancy serum, plasma and urine samples have been stored. Cord blood samples have been obtained and stored and DNA extraction has been completed on all cord and pregnancy samples. The cohort is broadly characteristic of the city's maternal population. Mean age of the mothers at study recruitment was 27 years old. Researchers are looking at the links between the circumstances of a child's birth, the context in which they grow up, their health and well-being and their educational progress.
For DNAm assays, 1000 mother-child pairs were selected (2000 individuals) from the participants who had completed a pregnancy OGTT (85% of BiB participants completed an OGTT), genome wide data on both mother and offspring (at the time of selection ~65% of those with OGTT) and were of either Pakistani or White British ethnic origin (the two largest homogeneous ethnic groups jointing reflecting 90% of the cohort). Within these criteria 500 Pakistani and 500 White British mother-child pairs were randomly selected.
Ethical approval for the data collection was granted by Bradford Research Ethics Committee (Ref 07/H1302/112). On registration with the study, pregnant mothers gave written informed consent for themselves and on behalf of their child.

DNAm assays

A total of 500 ng high molecular weight DNA was bisulfite-converted using the EZ-96 DNA methylation kit (Zymo Research, Orange, CA, USA). DNAm was quantified using Illumina HumanMethylation EPIC Arrays (Illumina, San Diego, CA, USA). During the data generation process, a wide range of batch variables were recorded in a purpose-built laboratory information management system (LIMS). Sample QC and normalization was performed using with *meffil*[9]. Samples failing QC (average probe $p > 0.01$) were excluded from further analysis. As an additional QC step genotype probes were compared with genotype data

from the same individual to identify and remove any sample mismatches. Furthermore, samples failed on control probes (bisulfite 1 and bisulfite 2) were also excluded from the analysis. Finally, 864 samples passed QC. Samples were normalized using functional normalization[10] using *meffil*[9]. Data was normalised using 7 control probe PCs derived from the technical probes informed by *meffil* scree plots. For the GoDMC analysis, DNAm sites that were overlapping with the Illumina Infinium HumanMethylation450 BeadChip array were included.

Genotyping assays

A total of 200ng high molecular weight DNA was genotyped using either the Infinium Human Core Exome-24 v1.1 arrays or the Infinium global screen-24+v1.0 arrays (Illumina, San Diego, CA, USA). Samples were pre-processed using GenomeStudio 2011.1 (genotyping). Samples with Call_Rate < 0.95 were excluded. Poorly performing SNPs were identified based on Call_Freq < 0.97, Cluster Sep ≤ 0.3, AB R Mean ≤ 0.2, BB R Mean ≤ 0.2, AA R Mean ≤ 0.2, 10% GC Score ≤ 0.2, MI Errors > 2 and Rep Errors > 2. A reference allele file was used to check the alleles from here:
ftp://ftp.ncbi.nih.gov/snp/organisms/archive/human_9606_b142_GRCh37p13/VCF/
Some variable names that were not rs# were switched to rs# from exm# using the list generated from below and swapping them in the original PLINK files for the directly genotyped data:
ftp://webdata:webdata@ussd-ftp.illumina.com/Downloads/ProductFiles/HumanCoreExome-24/Product_Support_Files/humancoreexome-24-v1-0-a-loci-name-to-rsid-conversion-file.zip
Using the reference lists, variables where the chromosome is not in the range 1-24 were removed and the labels for chromosomes 23 and 24 were renamed to X and Y. Those that were able to be flipped were flipped and those that could be correctly re-assigned the minor allele were switched. This process was repeated to catch those where the strand and the minor allele were not matching the reference. Multi-allelic SNPs were often discarded. This resulted in a VCF file containing all original 459,340 variables.
The VCF file above was submitted to the Sanger Imputation Service using the "UK10K + 1000 Genomes Phase 3" as a reference panel and "pre-phase with EAGLE2 and impute" as the pipeline. The 1000 Genomes Phase 3 panel was chosen as it contains samples originating from several global populations.

Data availability

Data are available to researchers who submit an expression of interest to the Born in Bradford Executive Group. We review applications monthly and aim to respond within 8 weeks. More details of data available and how to apply for access on the Born in Bradford website: https://borninbradford.nhs.uk/research/.

Acknowledgements

Born in Bradford is only possible because of the enthusiasm and commitment of the Children and Parents in BiB. We are grateful to all the participants, practitioners and researchers who have made Born in Bradford happen. 450K DNAm and genotype array data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Funding

BiB receives core funding from the Wellcome Trust (WT101597MA), the British Heart Foundation (CS/16/4/32482), a joint grant from the UK Medical Research Council (MRC) and UK Economic and Social Science Research Council (ESRC) (MR/N024397/1) and the National Institute for Health Research (NIHR) under its Collaboration for Applied Health Research and Care (CLAHRC) for Yorkshire and Humber. The research presented in this paper, including obtaining genome-wide and epigenome-wide DNAm data is supported by the US National Institute of Health (R01 DK10324) and European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 669545. NK and DAL work in a unit that receives support from the University of Bristol and UK MRC (MC_UU_00011/6) and DAL is an NIHR senior investigator (NF-SI-0611-10196).

Brisbane Systems Genomics Study (BSGS)

As described in detail elsewhere[23], individuals present in this study were recruited as part of the Brisbane Twin Nevus and cognition studies (known as BTN and MAPS respectively). Adolescent MZ and DZ twins, their siblings, and their parents were recruited over a 16 year period into an ongoing study of the genetic and environmental factors influencing pigmented nevi and the associated risk of developing skin cancer and cognition. The individuals are of northern European origin as confirmed by a Principal Component Analysis (PCA) comparing individuals in this study to HapMap3 populations. This study was approved by the Human Research Ethics Committee of the Queensland Institute for Medical Research. All participants gave informed written consent.

DNAm assays

DNAm was measured using Illumina HumanMethylation450 BeadChips as described in detail elsewhere[24]. Briefly, bisulfite converted DNA samples were hybridised to the 12 sample, Illumina HumanMethylation450 BeadChips using the Infinium HD Methylation protocol and Tecan robotics (Illumina, San Diego, CA, USA). DNAm probes were removed if they had a bead count less than 3 in >10% of samples, a probe detection p-value of >0.01 in >10% samples. All samples were confirmed to have concordant sex and genotypes across DNAm and SNP arrays. Following QC, a functional normalisation was performed using the package *meffil*[9]. Finally, 601 samples with 484973 probes were available for analysis.

Genotyping assays

All individuals were genotyped on Illumina 610-Quad Beadchip arrays. Full details of genotyping procedures are given elsewhere[25]. Standard QC filters were applied, leaving 528,509 SNPs. Data were imputed to 1000 Genomes Phase 3 reference panel[13] using the Michigan Imputation Server[26]. Imputed SNPs were filtered to have MAF > 0.01 and imputation quality score > 0.8, leave 7,704,504 SNPs that were converted to best guess binary plink format without a probability threshold.

Data availability

DNAm data for the BSGS is available at GEO under accession code GSE56105.

Acknowledgements

We gratefully acknowledge the participation of the twins and their families. We thank Marlene Grace, Ann Eldridge and Kerrie McAloney for sample collection and processing; the staff of the Molecular Epidemiology Laboratory at QIMR for DNA sample processing and preparation.

Funding

The Brisbane Systems Genetics Study (BSGS) was supported by NHMRC grants 1010374, 496667, 1046880. A.F.M., P.M.V., and G.W.M. are supported by the NHMRC Fellowship Scheme (1083656, 1078037 and 1078399) and grants (1050218). We acknowledge funding by the Australian Research Council (A7960034, A79906588, A79801419, DP0212016, DP0343921), and the Australian National Health and Medical Research Council (NHMRC) Medical Bioinformatics Genomics Proteomics Program (grant 389891) for building and maintaining the adolescent twin family resource through which samples were collected.

DNBC GOYA-offspring study

The Danish National Birth Cohort study (DNBC), the Genetics of Overweight Young Adults and their offspring (GOYA-offspring study) includes a subset of 91,387 pregnant women recruited to the Danish National Birth Cohort (DNBC) during 1996–2002 and their children[27]. A total of 67,853 women who had given birth to a live born infant, had provided a blood sample during pregnancy and had body mass index (BMI) information available, 3.6% of these women with the largest residuals from the regression of BMI on age and parity were selected for the GOYA study. The BMI for these 2451 women ranged from 32.6 to 64.4. From the remaining cohort, a random sample of 2450 mothers was also selected. In total, 3908 mothers were successfully genotyped. DNAm data were generated for the offspring of 1000 mothers in the GOYA study, equally distributed between “cases” with a BMI>32 and the random sample who were sampled from the remaining BMI distribution[28]. All women in the GOYA study provided written informed consent that their data and biological material could be used in scientific studies of health in women and children when entering the DNBC. The GOYA study was approved by the regional scientific ethics committee and by the Danish Data Protection Board.

DNAm assays

All data was imported into R version 3.2.0 and processed using *meffil*[9]. In total, there are 1010 samples belonging to 1000 children. Samples were extracted from cord blood. Ten samples were poor quality samples and were therefore repeated in the lab. Of the GOYA samples with DNAm profiles, 933 have been successfully genotyped. Samples were removed due to genotype mismatches between 65 genotypes extracted from the genotype and 65 SNP probes extracted from DNAm assays. Furthermore, DNAm quality was checked by: sex mismatches (23 samples), the median intensity methylated vs unmethylated signal for all control probes (n=8), bisulfate 1 probes (n=8), bisulfate II probes (n=2), dyebias (n=0), detection p-value (n=7), low bead numbers (n=0) and post normalization checks (n=5). The data was normalized using functional normalization[10] and 10 PCs were used to capture technical variation. After removing low quality sample samples, we had 892 samples

including 4 replicates, which gave us a final n of 888. GOYA was normalized using 15 control probe PCs based on *meffil* scree plots.

Genotyping assays

For the GOYA children cord blood was collected at birth. The children were genotyped using the Illumina Infinium HumanCoreExome Beadchip (Illumina, San Diego, CA, USA) and genotypes were made using the Genotyping module, version 1.9.4 of GenomeStudio software, version 2011.1 (Illumina, San Diego, CA, USA). The genotyping was carried out at the Novo Nordisk Center for Basic Metabolic Research, University of Copenhagen. We applied a >95% genotype call rate filter for the inclusion of SNPs. Additional genotypes were imputed into 1000 Genomes Phase 1[13] using IMPUTE 2[14].

Data availability

Data are available by request from DNBC, https://www.dnbc.dk/

Acknowledgments

The authors want to thank the many pregnant women who have taken part in the DNBC study. Without them, there would be no cohort. 450K DNAm array data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Funding

The Danish National Research Foundation has established the Danish Epidemiology Science Centre that initiated and created the Danish National Birth Cohort (DNBC). The cohort is furthermore a result of a major grant from this foundation. Additional support for the Danish National Birth Cohort is obtained from the Pharmacy Foundation, the Egmont Foundation, the March of Dimes Birth Defects Foundation, the Augustinus Foundation, and the Health Foundation. The DNBC biobank is a part of the Danish National Biobank resource, which is supported by the Novo Nordisk Foundation.
The GOYA study was conducted as part of the activities of the Danish Obesity Research Centre (DanORC, www.danorc.dk) and the MRC Integrative Epidemiology Unit, and genotyping was funded by the Wellcome Trust (WT 084762MA). Generation of DNAm data was supported by the MRC Integrative Epidemiology Unit which is supported by the Medical Research Council (grant number [MC_UU_12013/1-9]) and the University of Bristol. The genotyping in the GOYA-offspring has been financed by the Novo Nordisk Foundation Center for Basic Metabolic Research.

Dunedin Multidisciplinary Health and Development Study (Dunedin Study)

Participants are members of the Dunedin Multidisciplinary Health and Development Study, a longitudinal investigation of the health and behavior of a representative birth cohort of consecutive births between April 1972 and March 1973 in Dunedin, New Zealand. The cohort of 1,037 children (52% boys) was constituted at age 3 as 91% of eligible births resident in the province. The cohort represents the full range of socioeconomic status on

NZ's South Island and matches the NZ National Health and Nutrition Survey on adult health indicators (e.g., BMI, smoking, GP visits)[29]. Cohort members are primarily white; approximately 7% self-identify as having any non-white ancestry, matching the South Island. Follow-up assessments were conducted at ages 5, 7, 9, 11, 13, 15, 18, 21, 26, 32, and most recently 38, when 95% of the 1,007 living study members underwent assessment in 2010-2012. The study protocol was approved by the institutional ethical review boards of the participating universities. Study members gave informed consent before participating.

DNAm assays

DNAm data were generated using the Infinium HumanMethylation450 BeadChips (Illumina, CA, USA). Whole-blood was collected from the non-Maori participants in K_2EDTA vacutainer tubes (BD, NJ, USA) at ages 26 and 38. DNA was extracted from the buffy coat using standard procedures[29,30,31]. DNA Samples were arranged into 96-well plates so that within-individual age-26 and -38 DNA samples were hybridized in the same row of the arrays. (i.e. age 26 and 38 DNA samples from the same individual occupy array columns 1 and 2 of the same row). ~500ng of DNA from each sample was treated with sodium bisulfite using the EZ-96 DNA Methylation kit (Zymo Research, CA, USA). Array analysis was performed by the Duke University Molecular Physiology Institute Genomics Core Facility using the iScan platform (Illumina).

Data were processed and normalized using the *methylumi* (v2.14.0)[2] Bioconductor package from the R statistical programming environment, and subjected to QC analyses. Briefly, the method corrects Cy3 and Cy5 dye bias and recalculates the betas based on the corrected intensities against a reference array, which defaults to the first chip in the set. Samples were removed if the average detection p was >= 0.001. To confirm genetic identity of the DNA samples, we assessed genotype concordance between SNP probes on the 450K array and data generated using Illumina OmniExpress12v1.1 genotyping BeadChips. Principal component analysis (PCA) was performed independently for each normalized age-specific dataset and the first two components plotted. Samples formed two major clusters separating on the 1st component, which corresponded to recorded sex, and was thus used to confirm sex. After QC, blood-derived DNAm data was available for 895 individuals. Of these, 83% have valid DNAm data available from both assessment phases.

Genotyping assays

Whole genome SNP genotyping for the Dunedin Study was performed using OmniExpress12v1.1 BeadChips (Illumina). Dunedin DNA samples from 932 individuals were genotyped. Of these samples, 96.8% were derived from venous blood, the remainder from buccal swab. The majority of samples (88.7%) were collected at phase 38 assessment; the remainder were derived from phase 26 DNA since either a) a DNA sample was not obtained at phase 38 assessment or b) the phase 38 DNA sample was of insufficient yield or quality for analysis. Samples were arranged into 96-well plates that also included one duplicate within-plate Dunedin DNA sample and one individual DNA sample from a well-characterized CEPH family (NA7057, NA6990, NA6983 and NA6988; Coriell institute, NJ, USA). The duplicates and CEPH samples serve as within- and across- plate controls to check for genotyping inconsistencies. Genotyping was performed by Duke University Molecular Physiology Institute Genomics Core Facility using the iScan platform (Illumina).

Scan data were processed within GenomeStudio software (Illumina), and QC was performed following manufacturer's guidelines. Briefly, the QC steps undertaken were: 1) Poorly performing samples (sample call rate <98% upon initial clustering) were omitted, and samples re-clustered. 2) After re-clustering, individual SNPs were manually reviewed and edited or removed if any of the following criteria were met: More than three discrete genotype clusters present; Cluster Separation metric <0.3; SNP call frequency <95%. Heterozygote genotype cluster intensity (AB-R mean) <0.25; Heterozygote clusters shifted toward homozygote clusters (AB T mean <0.2 or >0.8); One or more parent-child (CEPH samples) and/or replication errors (CEPH and Dunedin duplicate samples); Over- or under-representation of heterozygotes, (Het Excess <-0.11 or >0.12); MAF <0.01. In addition, Y chromosome SNPs were removed if calls were made for known female samples, and X chromosome SNPs were removed if heterozygote calls were made for known male samples. After QC, genotype data were available for 98.5% of SNPs assayed, whilst 98.2% of Dunedin DNA samples yielded data at call rates greater than 98%. The data were output into PLINK format[11] for downstream data management and analysis.
We imputed additional SNPs using the IMPUTE2 software (version 2.3.1)[14] and 1000 Genomes version 3 reference panel[13]. Imputation was conducted on autosomal SNPs appearing in dbSNP (v140)[32] that were called in >98% of the Dunedin Study samples. Invariant SNPs were excluded. Pre-phasing and imputation were conducted using a 50Mbp sliding window. The resulting genotype database included genotyped SNPs and SNPs imputed with 90% likelihood of a specific genotype among the non-Maori members of the Dunedin cohort (n=918) and in HWE (p>0.01 for all).

Data availability

Data are available via a managed access system (contact: ac115@duke.edu). To promote awareness of the data among other scientists, we have done the following: 1) Registered the E-Risk and Dunedin Studies on the UK-MRC Cohort registry website, https://www.mrc.ac.uk/research/facilities-and-resources-for-researchers/cohort-directory/; 2) Registered the Dunedin Study with the Integrative Analysis of Longitudinal Studies of Aging and Dementia (IALSA) research network (NIH/NIA P01AG043362) https://www.maelstrom-research.org/mica/network/ialsa; 3) Registered the Dunedin Study on the UK-MRC JPND Global Cohort Portal, which recognizes the importance of longitudinal cohort studies for neurological disease research; http://www.neurodegenerationresearch.eu/initiatives/jpnd-alignment-actions/longitudinal-cohorts/; 4) Published[29]; (5) Posted all data-capture forms for every wave of the longitudinal study on the Dunedin Study website: http://dunedinstudy.otago.ac.nz/studies/dunedin-study-data-directories

Acknowledgements

We thank the Dunedin Study members, Unit research staff, and Study founder Phil Silva.

Funding

The Dunedin Multidisciplinary Health and Development Research Unit is supported by the New Zealand Health Research Council and New Zealand Ministry of Business, Innovation and Employment (MBIE). This research received support from the US-National Institute of Aging (grant number R01AG032282) and the UK Medical Research Council (grant number MR/P005918). Additional support was provided by the Jacobs Foundation. This work used a

high-performance computing facility partially supported by grant 2016-IDG-1013 ("HARDAC+: Reproducible HPC for Next-generation Genomics") from the North Carolina Biotechnology Center.

Environmental Risk Longitudinal Twin Study (E-Risk)

Participants were members of the Environmental Risk (E-Risk) Longitudinal Twin Study, which tracks the development of a 1994-95 birth cohort of 2,232 British children[33]. Briefly, the E-Risk sample was constructed in 1999-2000, when 1,116 families (93% of those eligible) with same-sex 5-year-old twins participated in home-visit assessments. This sample comprised 56% monozygotic (MZ) and 44% dizygotic (DZ) twin pairs; sex was evenly distributed within zygosity (49% male). The study sample represents the full range of socioeconomic conditions in Great Britain, as reflected in the families' distribution on a neighborhood-level socioeconomic index (ACORN [A Classification of Residential Neighbourhoods][33,34], developed by CACI Inc. for commercial use): 25.6% of E-Risk families live in "wealthy achiever" neighborhoods compared to 25.3% nationwide; 5.3% vs. 11.6% live in "urban prosperity" neighborhoods; 29.6% vs. 26.9% in "comfortably off" neighborhoods; 13.4% vs. 13.9% in "moderate means" neighborhoods; and 26.1% vs. 20.7% in "hard-pressed" neighborhoods. E-Risk underrepresents "urban prosperity" neighborhoods because such households are often childless.

Home visits were conducted when participants were aged 5, 7, 10, 12 and most recently, 18 years (93% participation). Our epigenetic study used DNA from a single tissue: whole blood. At age 18, whole blood was collected in 10mL K2EDTA tubes from 1,700 participants and DNA extracted from the buffy coat. Parents gave informed consent and twins gave assent between 5-12 years and then informed consent at age 18. The Joint South London and Maudsley and the Institute of Psychiatry Research Ethics Committee approved each phase of the study.

DNAm assays

We assayed 1669 blood samples (out of 1700); 31 samples were not useable (e.g., due to low DNA concentration). ~500ng of DNA from each sample was treated with sodium bisulfite using the EZ-96 DNA Methylation kit (Zymo Research, CA, USA). DNAm was quantified using the Illumina Infinium HumanMethylation450 BeadChip ("Illumina 450K array") run on an Illumina iScan System (Illumina, CA, USA) at the University of Exeter Medical School by the Complex Disease Epigenetics Group (as described in[35]). Twin pairs were randomly assigned to bisulfite-conversion plates and Illumina 450K arrays, with siblings processed in adjacent positions to minimize batch effects. Data were imported using the methylumIDAT function in *methylumi*[2] and subjected to QC analyses, checking for sex mismatches, genotype data that did not concur with those typed on Illumina OmniExpress24v1.1 arrays, and excluding low intensity samples. In total, samples from 1658 participants passed our QC pipeline and were included in the analyses presented here. Data were processed with the pfilter function from the *wateRmelon* package[4] excluding 0 samples with >1% of sites with a detection p >0.05, 567 sites with beadcount <3 in 5% of samples and 1448 probes with >1% of samples with detection p >0.05. The data were normalized with the dasen function from the *wateRmelon* package[4].

Genotyping assays

We used Illumina HumanOmni Express 24 BeadChip arrays (Version 1.1; Illumina) to assay common single-nucleotide polymorphism (SNP) variation in the genomes of cohort members. We imputed additional SNPs using the IMPUTE2 software (Version 2.3.1;[14] and the 1000 Genomes Phase 3 reference panel[13]. Imputation was conducted on autosomal SNPs appearing in dbSNP (Version 140[32]) that were “called” in more than 98% of the samples. Invariant SNPs were excluded. The E-Risk cohort contains monozygotic twins, who are genetically identical; we therefore empirically measured genotypes of one randomly-selected twin per pair and assigned these data to their monozygotic co-twin. Prephasing and imputation were conducted using a 50-million-base-pair sliding window. The resulting genotype database included genotyped SNPs and SNPs imputed with 90% probability of a specific genotype.

Data availability

These DNAm data are available in GEO under accession number GSE105018.

Acknowledgements

We thank the Study mothers, fathers, and their twins, the twins' teachers, Robert Plomin and his TEDS research team.

Funding

The E-Risk Study is funded by the Medical Research Council (G1002190). Additional support was provided by the National Institute of Child Health and Human Development (HD077482), a Distinguished Investigator Award from the American Asthma Foundation to Dr. Mill, and by the Jacobs Foundation. This work used a high-performance computing facility partially supported by grant 2016-IDG-1013 (“HARDAC+: Reproducible HPC for Next-generation Genomics") from the North Carolina Biotechnology Center.

Estonian Genome Center, University of Tartu (EGCUT CTG and EGCUT Asthma)

The EGCUT is a population based biobank which comprises health, genealogical and ‘omics’ data of more than 51,530 individuals ≥18 years of age, closely reflecting the age distribution in the adult Estonian population[36]. Participants of the EGCUT have been recruited by clinicians at their offices or data collectors at recruitment offices of the EGCUT. A computer assisted personal interview was completed for each participant, including personal data (place of birth, place(s) of living, nationality etc.), genealogical data (family history, three generations), educational and occupational history and lifestyle data (physical activity, dietary habits, smoking, alcohol consumption, quality of life). Anthropometric and physiological measurements were also recorded. All diseases are defined according to ICD10 coding. Women filled in an additional questionnaire relating to women’s health. The collection of blood samples and the data is conducted according to the Estonian Human Gene Research Act.
The samples for this study were selected from two cohorts with DNAm data: the Center for Translational Genomics (CTG) cohort and the asthma cohort. The EGCUT CTG cohort

consists of 1000 individuals who have been re-contacted for a second time-point sample. The ECGUT asthma cohort is a case-control set of individuals with early onset asthma, BMI<30, non-smokers and matched controls, selected for an epigenome wide association study of asthma. The study was approved by the Ethics Review Committee of Human Research of the University of Tartu, Estonia and it was carried out in compliance with the Helsinki Declaration. All of the participants were older than 18 and signed a broad informed consent.

DNAm assays

DNA from whole blood was extracted by the salting-out method using 10 M ammonium acetate. The DNA was precipitated in isopropanol, washed in 70% ethanol, and finally resuspended in 1X TE buffer. 500 ng of genomic DNA was treated with sodium bisulfite using the EZ DNA Methylation Kit (Zymo Research Corporation) according to the manufacturer's instructions. Bisulfite converted DNA was amplified, fragmented and hybridised to Illumina Infinium Human Methylation450 Beadchip using standard Illumina protocol. QC and normalization of the DNAm data was performed using R package *meffil* (with default parameters)[9]. Quality of DNAm data was evaluated by sex check, X-Y ratio, the median intensity methylated vs unmethylated signal for all control probes, dye bias, bisulfite conversion control probes, detection p-value, low bead numbers, genotype concordance, and post normalization checks. Finally, 305 (EGCUT CTG) and 105 (EGCUT Asthma) samples with genotype data available passed QC. EGCUT CTG and EGCUT Asthma were normalized using 8 control probe PCs derived from the technical probes informed by *meffil* scree plots.

Genotyping assays

DNA from the samples were genotyped using HumanOmniExpress BeadChips (Illumina) at the Estonian Genome Center, according to the manufacturer's instructions. Samples were excluded based on sample call-rate<0.95 as well as sample heterozygosity test, check for population outliers and samples with wrong/unidentifiable sex. SNPs with marker call rate<0.95 and HWE p< 1e-6 were excluded. The dataset was imputed using the 1000 Genomes project phase 3 reference[13] with IMPUTE v2[14]. SNPs with MAF>0.01 and imputation info score>0.8 were retained for the analyses.

Data availability

Data of the subjects from the Estonian biobank are handled in accordance with the regulations of the Human Genes Research Act. Data can be accessed upon ethical approval by submitting a data release request to the Estonian Genome Center, University of Tartu (http://www.geenivaramu.ee/en/access-biopank/data-access).

Acknowledgements

We thank Viljo Soo for his assistance with laboratory work, and Karit Mikkel and Mari-Liis Tammesoo for their assistance with data management. This work was carried out in part in the High Performance Computing Center of University of Tartu.

Funding

This work was supported by the Estonian Research Council grants IUT20-60, PRG184, IUT24-6: Estonian Centre for Genomics; the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.15-592 0012), and EU H2020 grant ePerMed (grant no. 692145).

European Prospective Investigation of Cancer - Norfolk (EPIC-Norfolk)

The EPIC-Norfolk study[37] is a prospective cohort study that recruited 25,639 individuals aged between 40-79 years at baseline in 1993-1997. The cohort was representative of the general population of England and Wales for age, sex, anthropometric measures, blood pressure and serum lipids, but differed in that 99.7% of the cohort were of European descent. We defined a random sub-cohort of the whole EPIC-Norfolk study population excluding known prevalent cases of diabetes at baseline using the same definitions as used in the InterAct project[38] who had available genotype data. Incident T2DM cases were ascertained from multiple sources: two follow-up health and lifestyle questionnaires providing self-reported information on doctor-diagnosed diabetes or medications; medications brought to the second clinical exam; and medical record linkage. Record linkage to external sources included the listing of any EPIC-Norfolk participant in the general practice diabetes register, local hospital diabetes register, hospital admissions data with screening for diabetes-related admissions, and Office of National Statistics mortality data with coding for diabetes. Participants who self-reported a history of diabetes which could not be confirmed against any other sources were not considered as confirmed cases. Follow-up was censored at date of diagnosis of T2DM, 31 July 2006, or date of death, whichever came first. All participants in the EPIC-Norfolk study gave signed informed consent and the study was approved by the Local Research Ethics Committee.

DNAm assays

DNAm was measured using the Illumina HumanMethylation450 array. Bisulfite conversion of DNA was performed using the EZ DNA methylation kit (Zymo Research, Orange, CA, USA). For 1,378 EPIC-Norfolk participants, DNAm was measured in DNA extracted from whole blood samples collected at baseline. Converted DNA was assayed by PCR (Polymerase Chain Reaction) and gel electrophoresis. Each 96 well DNA sample plate contained two duplicate samples. The average correlation between the duplicate samples was 98%.
In EPIC-Norfolk, epigenome-wide DNAm data were analysed in R (version 3.2.2). Initial QC was performed as recommended by the array manufacturer; DNAm intensity values were corrected using the Illumina Background Correction algorithm as implemented in *minfi*[39], DNAm intensities with a detection p ≥0·01 were set to 'missing' and DNAm intensity beta values were calculated for each DNAm marker per sample. For duplicate samples, the sample with the lower DNAm detection percentage was excluded.
Sample call rates were calculated as the proportion of missing data in each sample, by autosomal, X and Y chromosomes. For the autosomal data, 77 samples with a call rate ≤0·99 were excluded. All samples passed the call rate threshold on the X chromosome. For the Y chromosome, seven male samples that did not pass the call-rate and two further female samples were excluded. Distributions of DNAm intensities were also inspected by

autosomal and sex chromosomes, and separately in females and males leading to the exclusion of two additional samples that had an unusual distribution of DNAm intensities. Additionally we performed further QC using the *meffil*[9] package in R. Briefly DNAm quality was additionally checked for: the median intensity methylated vs unmethylated signal for all control probes (N=3), dyebias (N=0), low bead numbers (N=0), genotype concordance (N=5) and post normalization checks (N=0). We took forward the samples that were present in both the genotype and DNAm datasets with a final dataset of 1,105 for further analyses. Samples were normalized using functional normalization[10] using *meffil*[9]. Normalization was done on the DNAm data using 10 control probe PCs derived from the technical probes informed by *meffil* scree plots.

Genotyping assays

Genotyping was performed using the Axiom UKBiobank chip. Genotyping underwent a number of QC procedures including (a) routine quality checks carried out during the process of sample retrieval, DNA extraction, and genotype calling; (b) checks for genotype batch effects, plate effects, departures from Hardy-Weinberg equilibrium, sex effects, array effects, and discordance across control replicates; (c) individual and genetic variant call rate filters. Genotypes were filtered to include an MAF >1 % and imputation info score >0.8 and were imputed to 1000G phase3 (Dec 2014)[13] with phasing using SHAPEIT v2[12] and imputation using IMPUTE v2.3.1[14].

Data availability

For investigators wishing to access data from the EPIC-Norfolk please contact the study management committee http://www.srl.cam.ac.uk/epic/contact/

Acknowledgements

We are grateful to all of the participants and staff of the EPIC-Norfolk study.

Funding

EPIC-Norfolk is supported by programme grants from the Medical Research Council (MRC) [G9502233; G0401527; G100143] and Cancer Research UK [C864/A8257]. The generation and management of the Illumina 450K DNAm array data in this cohort is supported through the MRC Cambridge initiative in metabolomic science [MR/L00002/1]. The genome-wide genotyping data in EPIC-Norfolk was funded by an MRC award MC_PC_13048. This work is also supported by MRC programme grants [MC_UU_12015/1, and MC_UU_12015/2].

Italian cardiovascular section of EPIC (EPICOR Study)

The Italian cardiovascular section of EPIC (EPICOR study)[40] is a case-cohort study nested in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Italy cohort. The EPIC-Italy cohort comprises about 50,000 participants[40,41] enrolled between 1992 and 1998, who provided at enrollment a detailed dietary and lifestyle questionnaire and a blood sample that was stored in liquid nitrogen for later use. The EPIC cohort is regularly followed up for the occurrence of cancers and other non-communicable diseases of adulthood. Four EPIC-Italy centers (Turin, Varese, Naples, and Ragusa) provided samples to EPICOR. The whole

EPICOR study comprises more than 1,500 subjects with cardiovascular outcomes such as myocardial infarction (MI), acute coronary syndrome, ischemic cardiomyopathy, coronary or carotid revascularization, ischemic- or hemorrhagic stroke. Within the EPICOR cohort, a subset of 584 subjects (292 MI cases and 292 matched controls) was analyzed as a nested case-control study and underwent DNAm analysis[42] and whole genome genotyping. All volunteers signed an informed consent form at enrollment in the respective studies. EPICOR study complies with the Declaration of Helsinki principles and conforms to ethical requirements. The EPIC study protocol was approved by Ethics Committees of the International Agency for Research on Cancer (Lyon, France), as well as by local Ethical Committees of the participant centers. The EPICOR study was approved by the Ethical Committee of the Italian Institute for Genomic Medicine (IIGM, formerly Human Genetics Foundation-Torino, HuGeF, Turin, Italy).

DNAm assays

DNAm was measured in DNA from WBCs collected at subject enrollment into EPIC and stored in liquid nitrogen. Genomic DNA was extracted from 400ul buffy coat from whole blood stored in liquid nitrogen at sample recruitment by an automated on-column DNA purification method (QIAsymphony instrument and QIAsymphony DNA Kits, QIAGEN GmbH, Germany), according to manufacturer's standard protocols. DNA integrity was checked by an electrophoretic run in standard TBE 0.5X buffer on a 1% low melting agarose gel (Sigma-Aldrich GmbH, Germany); DNA purity and concentration were assessed by a NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific Inc.). Five hundreds of genomic DNA were bisulfite converted (EZ-96 DNA Methylation-Gold Kit, Zymo Research Corporation) according to manufacturer's protocol. The DNAm status of more than 485,000 individual DNAm loci at a genome-wide resolution, was assessed by the Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, CA, USA) according to standard manufacturer's protocols. QC[9] and functional normalization[10] was performed: a total of 292 matched case-control pairs (584 subjects) and 484,683 sites passed QC and were retained for further analyses.

Genotyping assays

Whole-genome genotyping of 576 subjects was performed on the Infinium OmniExpressExome BeadChip (Illumina Inc., San Diego, CA, USA) according to standard manufacturer protocols.
Genotypes were assessed by GenomeStudio V2011.1(Illumina Inc., San Diego, CA).
Data QC procedures were carried out with PLINK (v1.07)[11] with the following criteria: Individual samples were tested for missingness (<0.05), heterozygosity and genetic sex discrepancies. We also tested for individual IBS (pi > 0.05), PCs were used to identify and remove ethnic outliers, and to adjust for population stratification in the downstream analyses. SNPs with minor allele frequency (MAF) < 1%, genotype call rate <95% and/or a HWE with P<1e−6 were excluded, together with individuals with >5% missing genotypes. Samples that perform poorly in the QC are excluded from further analysis.
After QCs, 7680624 SNPs and 556 samples were retained or further analyses.

Data availability

Requests for the data accession may be sent to Prof. Giuseppe Matullo (giuseppe.matullo@unito.it).

Acknowledgements

We wish to thank all the volunteers who participated in EPIC, and all the EPIC-Italy PI's who contributed samples for the EPICOR study (Dr. Vittorio Krogh, Dr. Paolo Vineis, Dr. Salvatore Panico, Dr. Rosario Tumino) for their continuous effort in managing and following up the cohort.

Funding

The work was supported by the Compagnia di San Paolo for the EPIC-Italy and EPICOR projects, the Italian Institute for Genomic Medicine (IIGM, formerly Human Genetics Foundation-Torino, HuGeF, Turin, Italy) and the MIUR ex60% grant. EPIC-Italy is further supported by a grant from the “Associazione Italiana per la Ricerca sul Cancro” (AIRC, Milan). This research received funding specifically dedicated to the Department of Medical Sciences from Italian Ministry for Education, University and Research (MIUR) under the programme "Dipartimenti di Eccellenza 2018 - 2022”.

Finnish Twin Cohort (FTC)

The Finnish twin cohort (FTC) of twins born before 1958 was established in year 1974 with the aim to examine the genetic, environmental, and psychosocial determinants affecting public health outcomes including several chronic diseases and health behaviours (www.twinstudy.helsinki.fi)[43]. Later cohorts of twins born in 1975-1979 (Finntwin16 study) and twins born in 1983-1987 (Finntwin12 study) have been studied. All cohorts have been assessed repeatedly and followed through multiple medical and population registries. Based on questionnaire response, subsets of pairs have been invited to in-person studies for detailed phenotyping and biosampling. Written informed consent, according to the current edition of the Declaration of Helsinki, was obtained from all subjects who were interviewed and/or gave DNA samples before the beginning of the studies. The collection of blood samples followed the recommendations given in the Declaration of Helsinki and its amendments. Data collection has been approved by the hospital district of Helsinki and Uusimaa, the ethics committee for epidemiology and public health (HUS-113-E3-01, HUS-346-E0-05, HUS 136/E3/01).

DNAm assays

DNA was extracted from whole blood using QIAamp DNA Mini kit (QIAGEN Nordic, Sollentuna, Sweden). Bisulfite conversion of DNA was completed using EZ-96 DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA) according to the manufacturer’s instructions, and the co-twins were always converted on the same plate to minimize potential batch effects. DNAm status was assessed using the Infinium HumanMethylation450 BeadChip at the Technology Centre, FIMM, University of Helsinki, Finland, The Microarray Consortium, Oslo, Norway, The Genomics Facility, University of Chicago, Chicago, IL, USA, and at The SNP&SEQ Technology Platform, University of Uppsala, Sweden. Sample QC

and normalization was performed using default parameters of *meffil*[9] R package version 3.2.0. The data was filtered to remove probes and samples with detection p>0.01 and that failed the threshold of 3 beads. 485378 probes from 1177 samples were available for the further analyses.

Genotyping assays

Genotyping was performed with the Human670-QuadCustom Illumina BeadChip at the Wellcome Trust Sanger Institute, and with the Illumina Human Core Exome BeadChip at the Wellcome Trust Sanger Institute and at the Broad Institute of MIT and Harvard. Standard post genotyping QC thresholds were applied for SNPs (MAF <0.01, SNP call rate <0.95, and HWE p<1e-6). Further, subjects with a call rate <0.95 were excluded, and a sample heterozygosity test, as well as sex and Multidimensional Scaling (MDS) outlier checks were done. Pre-phasing of the data was done with SHAPEIT2[12] and imputation with IMPUTE2[14] using the 1000 Genomes Phase III reference panel[13]. For data generated with the Human670-QuadCustom Illumina BeadChip the following post-imputation exclusion criteria were applied for SNPs: MAF<0.01, SNP call rate <0.95 (<0.99 for SNPs with MAF<0.05), HWE p<1e-6, and imputation info <0.4. For data generated with the Illumina Human Core Exome BeadChip the SNP exclusion criteria were otherwise identical, except that a threshold of minor allele count <2 was applied instead of a MAF cut-off. Further, the same sample quality thresholds as in post-genotyping QC were applied. QC and imputation for all Finnish GWAS data were done centrally at the Institute for Molecular Medicine FIMM, University of Helsinki, Helsinki, Finland.

Data availability

FTC data must obtain approval from the Data Access Committee of the Institute for Molecular Medicine Finland FIMM (fimm-dac@helsinki.fi). Note that anonymized individual level data can only be released after study has been approved by Research Ethics Committee of University of Helsinki and must be carried out in collaboration with FTC investigators. To ensure protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will depend on the nature of the requested data. For further information please contact Jaakko Kaprio (jaakko.kaprio@helsinki.fi).

Acknowledgements

We warmly thank the participating twin pairs and their family members for their contribution. Anja Häppölä, Mia Urjansson and Kauko Heikkilä are acknowledged for their valuable contribution in recruitment, data collection, and data management.

Funding

Phenotyping, genotyping, and epigenotyping of the Finnish twin cohorts has been supported by the Academy of Finland Center of Excellence in Complex Disease Genetics grants 213506, 129680; the Academy of Finland grants 100499, 205585, 118555, 141054, 265240, 263278 and 264146 (to Prof. Jaakko Kaprio), and 297908 and 251316 (to Dr. Miina Ollikainen), Sigrid Juselius Foundation (to Prof. Jaakko Kaprio and Dr Miina Ollikainen); Helsinki University Research Grants (to dr. Miina Ollikainen); EPITRAIN - FP7-PEOPLE-

2012-ITN, Grant Agreement 316758 funded by the European Union's Seventh Framework Programme.

Generation R Study

The Generation R Study is a prospective population-based cohort[44]. All pregnant women living in the city of Rotterdam, the Netherlands with a delivery date between April 2002 and January 2006 were invited to participate. In total, 9,778 mothers were enrolled in the study[44]. Written informed consent was obtained for all participating children. The Generation R Study was approved by the Medical Ethics Committee of Erasmus MC, University Medical Center Rotterdam, the Netherlands.

DNAm assays

Preparation and normalization of the HumanMethylation450 BeadChip array data was performed according to the CPACOR workflow[45] using the software package R. In detail, the idat files were read using the *minfi* package[39]. Probes that had a detection p above background (based on sum of methylated and unmethylated intensity values) ≥ 1e-16 were set to missing per array. Next, the intensity values were stratified by autosomal and non-autosomal probes and quantile normalized for each of the six probe type categories separately: type II red/green, type I methylated red/green and type I unmethylated red/green. Beta values were calculated as proportion of methylated intensity value on the sum of methylated+unmethylated+100 intensities. Arrays with observed technical problems such as failed bisulfite conversion, hybridization or extension, as well as arrays with a mismatch between sex of the proband and sex determined by the chr X and Y probe intensities were removed from subsequent analyses. Additionally, only arrays with a call rate > 95% per sample were processed further.

Genotyping assays

Cord blood for DNA isolation was available in 58% of all live-born children (N=5732). Sex-mismatch rate between genome-based sex and recorded sex was low (<0.5%), indicating extremely low possible contamination of maternal DNA. Missing cord blood samples were mainly due to logistical constraints at the delivery. Genome-wide association scans (GWAs) were run on the Illumina 610 Quad and 660 platforms. imputations to HapMap, 1000G, UK10K and HRC are available. Before imputation, SNPs were excluded if they had high levels of missing data (SNP call rate <98%), strong departures from HWE (p <1e-6), or low MAF (<0.1%). Imputations were performed using the MACH/Minimac software[18].

Data availability

Requests for data access are evaluated by the Generation R Management Team.

Acknowledgements

The Generation R Study is conducted by the Erasmus Medical Center in close collaboration with the School of Law and Faculty of Social Sciences of the Erasmus University Rotterdam, the Municipal Health Service Rotterdam area, Rotterdam, the Rotterdam Homecare Foundation, Rotterdam and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond

(STAR-MDC), Rotterdam. We gratefully acknowledge the contribution of children and parents, general practitioners, hospitals, midwives and pharmacies in Rotterdam. The study protocol was approved by the Medical Ethical Committee of the Erasmus Medical Centre, Rotterdam. Written informed consent was obtained for all participants. The generation and management of the Illumina 450K DNAm array data (EWAS data) for the Generation R Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Mr. Michael Verbiest, Ms. Mila Jhamai, Ms. Sarah Higgins, Mr. Marijn Verkerk and Dr. Lisette Stolk for their help in creating the EWAS database. We thank Dr. A.Teumer for his work on the QC and normalization scripts.

Funding

The general design of the Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development and the Ministry of Health, Welfare and Sport. The EWAS data was funded by a grant to VWJ from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810), by funds from the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by a grant from the National Institute of Child and Human Development (R01HD068437). This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements no.633595 (DynaHEALTH) and 733206 (LifeCycle). JFF has received funding from the European Joint Programming Initiative "A Healthy Diet for a Healthy Life" (JPI HDHL, NutriPROGRAM project, ZonMw the Netherlands no.529051022).

Glycyrrhizin in Licorice (GLAKU)

The adolescents of the Glaku (Glycyrrhizin in Licorice) cohort came from an urban community-based cohort comprising 1049 infants born between March and November 1998 in Helsinki, Finland[46]. In 2009–2011, initial cohort members who had given permission to be contacted and whose addresses were traceable (N = 920, 87.7% of the original cohort in 1998) were invited to a follow-up, of which 692 (75.2%) could be contacted by phone (mothers of the adolescents). Of them, 451 (65.2% of those who could be contacted by phone, 49% of the invited) participated in a follow-up at a mean age of 12.3 years (SD = 0.5, range 11.0–13.2 years).

DNAm assays

Following DNA extraction, genomic DNA was bisulfite modified using an EZ DNA methylation kit (Zymo Research, Orange, CA, USA). The protocol was as described by the manufacturer. Genome-wide DNAm was measured using the Illumina HumanInfinium Methylation BeadChip (Illumina, San Diego, CA, USA) following the manufacturer's protocol. Sample QC and normalisation was performed using *meffil*[9] in R.

Genotyping assays

DNA was extracted from blood samples (N=80) and saliva samples (N=277) donated at the 2009-2011 follow-up and genotyping was performed with the Illumina OmniExpress Exome

1.2 bead chip at the Tartu University, Estonia in September 2014 according to the standard protocols. Genomic coverage was extended by imputation using the 1000 Genomes Phase I integrated variant set (v3 / April 2012; NCBI build 37 / hg19) as the reference sample and IMPUTE2 software. Before imputing the following QC filters were applied: SNP clustering probability for each genotype > 95%, Call rate > 95% individuals and markers (99% for markers with MAF < 5%), MAF > 1%, HWE p > 1e-6. Moreover, heterozygosity, gender check and relatedness checks were performed and any discrepancies removed (N=2).

Data availability

Any interested researchers can obtain a de-identified dataset after having obtained an approval from the GLAKU Study Board. Data requests may be subject to further review by the national register authority and ethical committees. Any requests for data use should be addressed to GLAKU Study individual researchers.

Funding

This work was supported by the Academy of Finland, Hope and Optimism Initiative, the Signe and Ane Gyllenberg Foundation, the Emil Aaltonen Foundation, the Foundation for Pediatric Research, the Foundation for Cardiovascular Research, the Juho Vainio Foundation, the Sigrid Jusélius Foundation, the Yrjö Jahnsson Foundation, and the University of Helsinki Research Funds.

Acknowledgements

We thank all the GLAKU children and their parents for their enthusiastic participation. We also thank all the research nurses, research assistants, and laboratory personnel involved in the Predo study.

GSK study

Patients with recurrent unipolar depression were recruited from in- and out-patients at the Max Planck Institute of Psychiatry in Munich and psychiatric hospitals in Augsburg and Ingolstadt, located close to Munich. Each hospital contributed one-third of the patients. Patients were diagnosed by WHO-certified raters according to DSM-IV using the Schedule for Clinical Assessment in Neuropsychiatry. Only Caucasian patients over 18 years with at least two moderate-to-severe depressive episodes were included. Exclusion criteria were the presence of manic or hypomanic episodes, mood incongruent psychotic symptoms, the presence of a lifetime diagnosis of intravenous drug abuse and depressive symptoms only secondary to alcohol or substance abuse or dependence or to a medical illness or medication. Ethnicity was recorded using a self-report sheet for perceived nationality, first language and ethnicity of the subject himself, parents and all four grandparents. Controls matched for ethnicity (using the same questionnaire as for patients), sex and age (to 5-year intervals) were recruited at the Max Planck Institute of Psychiatry. Controls were randomly selected from a Munich-based community sample and screened for the presence of anxiety and affective disorders. All included controls were Caucasian and 93.04% were of German origin. These subjects thus represent a group of healthy individuals with regard to depression and anxiety.

The study was approved by the Ethics Committee of the Ludwig Maximilians University in Munich, Germany, and written informed consent was obtained from all subjects.

DNAm assays

Following DNA extraction, genomic DNA was bisulfite modified using an EZ DNA methylation kit (Zymo Research, Orange, CA, USA). The protocol was as described by the manufacturer. Genome-wide DNAm was measured using the Illumina HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA) following the manufacturer's protocol. Sample QC and normalisation was performed using *minfi*[39] in R. Briefly, DNAm quality was evaluated by: sex detection outliers, the median intensity methylated vs unmethylated signal and DNAm detection p. Functional normalization[10], followed by batch correction in Combat[47] was used.

Genotyping assays

On enrolment in the study, EDTA blood was drawn from each patient and each healthy control. DNA was extracted from fresh blood using the Puregene® whole blood DNA-extraction kit (Gentra Systems Inc., MN, USA). All individuals were genotyped using the Illumina 550k genotyping array.

Data availability

DNAm data is available in GEO (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE125105.

Funding

This work has been funded by the Bavarian Ministry of Commerce and by the Federal Ministry of Education and Research in the framework of the National Genome Research Network, Förderkennzeichen 01GS0481 and the Bavarian Ministry of Commerce.

INMA—INfancia y Medio Ambiente—(Environment and Childhood)

Mothers were enrolled at week 12 of pregnancy from 1997 to 2008 in seven regions of Spain (Flix, Granada, Menorca, Asturias, Gipuzkoa, Sabadell and Valencia). The cohort consisted of 3,768 children at birth. During the follow-up visits information on environmental exposures and health outcomes (reproductive, growth and obesity, lung function, allergies and neurovelopment) were assessed through questionnaires, biomarker measurements, clinical data, and physical exploration. The study website contains details of the design and data available in INMA project (http://www.proyectoinma.org/)[48]. The study was approved by the Ethical Committees of each participating centre and written consent was obtained from parents.

DNAm assays

DNAm data assessed with the Infinium HumanMethylation450 Beadchip is available in Sabadell subcohort for 391 whole blood samples collected at birth. Cord blood was extracted using the Chemagen kit (Perkin Elmer) at Spanish National Genotyping Centre (CEGEN).

DNA concentration was determined by a NanoDrop spectrophotometer (Thermo Scientific) and with the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies).
Blood DNAm data was produced in two laboratories: The Genome Analysis Facility of the University Medical Center Groningen (UMCG) in Holland as part of the MeDALL project and the Bellvitge Biomedical Research Institute (IDIBELL) in Barcelona as part of the BREATHE project. Both laboratories randomized the samples in batches and followed the Illumina protocol for the Infinium HumanMethylation450 BeadChip. Briefly, 500 ng of DNA were bisulfite-converted using the EZ 96-DNA methylation kit, and DNAm was measured through hybridization on the BeadChips. BeadChips were scanned with an Illumina iScan and image data was uploaded into the DNAm Module of Illumina's analysis software GenomeStudio and converted in β-values. Statistical analyses were adjusted for laboratory to control for batch effects.

Genotyping assays

DNA was obtained from cord blood or whole blood collected at 4y. 1071 children whose parents reported to be white and to be born in Spain or in European countries and that were not lost during the follow-up were selected for genotyping. Genome-wide genotyping was performed using the HumanOmni1-Quad Beadchip (Illumina) at CEGEN (Spanish National Genotyping Center). Genotype calling was done using the GeneTrain2.0 algorithm based on HapMap clusters implemented in the GenomeStudio software. PLINK was used for the genetic data QC[11]. We applied the following initial QC thresholds: sample call rate>98% and/or LRR SD<0.3. Then, we checked sex, relatedness, heterozygosity and population stratification. Genetic variants were filtered for SNP call rate>95%, MAF>1% and HWE p value>1e-6. Genetic imputation was performed using IMPUTE v2[14] and the cosmopolitan panel from 1000G[13] from release March 2012 as a reference.
The final genetic data set consisted of 396 children from INMA Sabadell. The overlap of children with genetic and epigenetic data is 351.

Data availability

Data are available by request from the INfancia y Medio Ambiente Executive Committee for researchers who meet the criteria for access to confidential data.

Acknowledgements

INMA researchers would like to thank all the participants for their generous collaboration. A full roster of the INMA Project Investigators can be found at https://www.proyectoinma.org/proyecto-inma/organizacion/.

Funding

This project was funded by grants from Instituto de Salud Carlos III (Red INMA G03/176, CB06/02/0041; PI041436; PI081151 incl. FEDER funds, PS09/00432), Generalitat de Catalunya-CIRIT 1999SGR 00241, Fundació La marató de TV3 (090430), CIBERESP, EU Commission (261357, ERC: 268479). CR-A was supported by a FI fellowship from Catalan Government (#016FI_B 00272).

Isle of Wight Third Generation (IOW F2)

The recruitment of newborns started in April 2010. Data used in the analyses were from infants born between April 2010 to May 2014[49]. In total, 200 newborns were recruited such that at least one of their parents is in the IOW 1989 birth cohort (IOW F1). For each infant, along with other phenotypic information such as gender and birth weight, status of wheezing and eczema was recorded, measures of wheal size from skin prick test as well as IgE were recorded.

DNAm assays

For 123 subjects, DNAm was measured in DNA extracted from cord blood using a simple salting out procedure[50]. Bisulfite-convertion was undertaken using the EZ 96-DNA methylation kit (Zymo Research, Irvine, CA, USA), following the manufacturer's standard protocol and DNAm measured using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA). All data was processed using *meffil*[9]. In total, 111 samples out of 123 passed QC and further normalised using *meffil*[9].

Genotyping assays

Whole-genome genotyping was performed on the Illumina Infinium OmniExpressExome-8 Kit (Illumina Inc., San Diego, CA, USA) according to standard manufacturer protocols. Genotypes were assessed by GenomeStudio V2011.1 (Illumina Inc., San Diego, CA). Data QCl procedures were carried out with PLINK (v1.07)[11] with the following criteria: Individual samples were tested for missingness (<0.05), heterozygosity and genetic sex discrepancies. SNPs with MAF < 1%, genotype call rate <95% and/or a HWE with P<1e−6 were excluded, together with individuals with >5% missing genotypes. We carried out imputation to 1000G phase1 v3[13] using SHAPEIT[12] and IMPUTE v2[14]. After QC, 8015356 SNPs and 111 samples were retained or further analyses.

Data availability

Data are available by request from Isle of Third Generation Study, (http://www.allergyresearch.org.uk/contact-us/).

Acknowledgments

We would like to thank all the participants of the Isle of Wight birth cohort, the research team at David Hide Asthma & Allergy Research Centre (Isle of Wight) for collecting the data, Nikki Graham for technical support and other members of the IoW research group for valuable discussion. DNAm data was generated by the Oxford Genomics Centre at the Wellcome Trust Centre for Human Genetics.

Funding

The IoW third generation cohort was funded by NIAID/NIH R01AI091905. This study was supported in part by NIAID/NIH R01AI091905 and R01AI121226.

Leiden Longevity Study

The Leiden Longevity Study (LLS) has been described in detail previously[51]. It is family-based study consists of 1671 offspring of 421 nonagenarian's sibling pairs of Dutch descent, and their 744 partners. In the current EWAS initiative 793 individuals with available DNAm data and phenotypic data (mean age = 60 years) were included.

DNAm assays

The generation of genome-wide DNAm data has previously been described[52]. Briefly, 500 ng of genomic DNA was bisulfite modified using the EZ DNA Methylation kit (Zymo Research) and hybridized on Illumina 450K arrays according to the manufacturer's protocols. Data were generated by the Human Genotyping facility (HugeF) of ErasmusMC, The Netherlands. Sample QC and normalization (10 PCs) were completed using *meffil*[9].

Genotyping assays

Details on the genotyping and QC methods have previously been detailed[53]. The genotype data were harmonized towards the 1000G reference[13] using Genotype Harmonizer and subsequently imputed using IMPUTE2[14]. Genotypes were filtered to have MAF > 0.01 and imputation quality score > 0.8.

Data availability

DNAm data were submitted to the European Genome-phenome Archive (EGA) under accession EGAS00001001077.

Acknowledgements

We thank all participants of the Leiden Longevity Study.

Funding

This study received funding from the European Union's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 259679, from the Innovation-Oriented Research Program on Genomics (SenterNovem IGE05007), the Centre for Medical Systems Biology, from the Netherlands Consortium for Healthy Ageing (grant 050-060-810), and from the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007), all in the framework of the Netherlands Genomics Initiative, Netherlands Organization for Scientific Research (NWO).

The Lothian Birth Cohorts of 1921 and 1936 (LBC1921 and LBC1936)

The Lothian Birth Cohorts of 1921 and 1936 stem from the Scottish Mental Surveys (SMS) of 1932 and 1947, respectively, when nearly all 11-year old children across Scotland (born in 1921 and 1936) completed a test of general intelligence (Moray House Test No. 12)[54]. Survivors from the SMS living in the Lothian area – Edinburgh and surrounding region –

were recontacted in later life and invited to participate in longitudinal studies of ageing: LBC1921 and LBC1936. The study members were first recontacted at mean ages 70 years (LBC1936) and 79 years (LBC1921), when they completed a series of cognitive, personality, medical, and health questionnaires. Blood samples were donated for the collection of biomarker, genetic, and epigenetic data. Longitudinal questionnaire and omics data are available in both cohorts: ages 70, 73, 76, 79 years in LBC1936; ages 79, 83, 87, 90, 92 in LBC1921. Following written informed consent, venesected whole blood was collected for DNA extraction in both LBC1921 and LBC1936. Ethics permission for the LBC1921 was obtained from the Lothian Research Ethics Committee (Wave 1: LREC/1998/4/183). Ethics permission for the LBC1936 was obtained from the Multi-Centre Research Ethics Committee for Scotland (Wave 1: MREC/01/0/56), the Lothian Research Ethics Committee (Wave 1: LREC/2003/2/29).

DNAm assays

Blood-based DNAm were obtained using the Illumina 450k DNAm array in both LBC studies. Blood samples were taken at mean ages 70 and 79 years in LBC1936 and LBC1921, respectively. QC information has been previously reported[55]. DNAm probes were removed if they had a bead count <3 in >10% of samples, a probe detection p-value of >0.01 in >10% samples. All samples were confirmed to have concordant sex and genotypes across DNAm and SNP arrays. Following QC, a functional normalisation was performed using the package *meffil*[9]. After these steps there were 477057/470707 probes available for analysis in 435/905 participants from LBC1921/LBC1936.

Genotyping assays

Genotyped data were generated at the University of Edinburgh Clinical Research Facility using the Illumina 610-Quadc1 array (San Diego, CA, USA). QC details have been reported previously[56]. Briefly, data preparation and QC steps included exclusions based on relatedness, sex discrepancies, low SNP call rate, evidence of non-European descent. SNPs with a MAF >1% and a Hardy-Weinberg pvalue >= 0.001. Data were imputed to 1000 Genomes Phase 3 reference panel using the Michigan Imputation Server[26]. Imputed SNPs were filtered to have MAF > 0.01 and imputation quality score > 0.8, leaving 6,920,489/7,189,799 SNPs rom LBC1921/LBC1936 that were converted to best guess binary plink format without a probability threshold.

Data availability

LBC1921 and LBC36 data are available on request from the Lothian Birth Cohort Study, Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh (email: I.Deary@ed.ac.uk). LBC data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

Acknowledgements

The authors thank all LBC1921 and LBC1936 study participants and research team members who have contributed, and continue to contribute, to ongoing LBC studies.

Funding

Phenotype collection in the Lothian Birth Cohort 1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society and The Chief Scientist Office of the Scottish Government. Phenotype collection in the Lothian Birth Cohort 1936 was supported by Age UK (The Disconnected Mind project) and the Medical Research Council (MR/M01311/1). DNAm typing was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. This work was conducted in the Centre for Cognitive Ageing and Cognitive Epidemiology, which is supported by the Medical Research Council and Biotechnology and Biological Sciences Research Council (MR/K026992/1), and which supports IJD. This research was supported by Australian National Health and Medical Research Council (grants 1010374, 1046880, and 1113400) and by the Australian Research Council (DP160102400). PMV, NRW, and AFM are supported by the NHMRC Fellowship Scheme (1078037, 1078901, and 1083656).

MARseille THrombosis Association study (MARTHA)

The MARTHA study is a collection of 1,592 patients with venous thrombosis (VT) recruited from the Thrombophilia centre of La Timone hospital (Marseille, France)[57,58,59,60]. All subjects had a documented history of VT, were free of chronic conditions, and were free of inherited thrombophilia including: anti-thrombin, protein C and protein S deficiencies and homozygosity for the Factor V Leiden and Factor II G20210A mutations. For the current project, 349 MARTHA patients were randomly selected for DNAm analysis.
Ethics approval was obtained from the "Département santé de la direction générale de la recherche et de l'innovation du ministère" (Projects DC: 2008-880 & 09.576). All subjects provided written informed consent in accordance with the Declaration of Helsinki.

DNAm assays

Genomic DNA was isolated from peripheral blood cells using an adaptation of the method proposed by[50]. For each sample, 1mg genomic DNA was bisulfite converted using the QiagenEpiTect 96 Bisulfite Kit. Then, 200ng of bisulfite-converted DNA at 50ng/ml was independently amplified, labelled, and hybridized to Infinium HumanMethylation450 BeadChip microarrays[61] and scanned with default settings using the Illumina iScan. All samples were processed at The Center for Applied Genomics (TCAG, Toronto, Canada). DNAm data were expressed as a β-value, a continuous variable over the [0-1] interval, representing the percentage of DNAm of a given site[62]. DNAm values were corrected for background by use of the Noob method implemented in the "*methylumi*" package[2], for dye bias following the manufacturer's recommendation (https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2011-1/genomestudio-methylation-v1-8-user-guide-11319130-b.pdf) and normalized for design type bias according to the SWAN method[63] implemented in the *minfi* R package[39]. QC and normalization were done simultaneously on the MARTHA and F5L-families datasets. Any probe with a detection p-value (as described in the "*minfi*" package) greater than 0.05 in more than 5% of the total processed samples was excluded from further

analyses. Probes that were cross-reactive or polymorphic at a targeted DNAm site[64,65] were also excluded.

Genotyping assays

MARTHA patients were genotyped with the Illumina Human 610 / 660W-Quad beadchips[57]. From the original sample of 1592 MARTHA patients, a subsample of 1011 VT patients were typed with the Illumina Human 610-Quad Beadchip while the remaining 586 VT patients were typed with the Illumina Human660W-Quad Beadchip.
Individuals with genotyping success lower than 95% (n=18) were excluded from the analyses as were individuals demonstrating close relatedness (n=67). This latter was assessed by pairwise clustering of identity by state distance (IBS) and multi-dimensional scaling (MDS) using the PLINK software[11]. The Eigenstrat program was further used to detect individuals of non-European ancestry. Autosomal SNPs that satisfied QC criteria (n=481,002)[66] were then used for imputing SNPs from the 1000 Genomes 2012-02-14 release reference dataset. Imputation was performed by use of MACH (v1.0.18.c) software[67]. All SNPs with acceptable imputation quality (r^2> 0.3)[68] and MAF > 0.01 were kept for further analyses.

Data availability

Illumina HumanMethylation450 array from MARTHA participants are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3127.

Funding

MARTHA data analyses were performed using the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05). The MARTHA Human450Methylation epityping was partially funded by the Canadian Institutes of Health Research (grant MOP 86466) and by the Heart and Stroke Foundation of Canada (grant T6484). The MARTHA project was also partially supported by the GENMED Laboratory of Excellence on Medical Genomics (ANR-10-LABX-0013) and the French Clinical Research Infrastructure Network on Venous Thrombo-Embolism (F-CRIN INNOVTE), two research programs managed by the National Research Agency (ANR) as part of the French Investment for the Future initiative

Netherlands Twin Register (NTR)

The subjects take part in longitudinal studies with the Netherlands Twin Register (NTR)[69,70] including the NTR biobank project between 2004 and 2011[71]. The NTR is a longitudinal twin-family study with no other selection criteria than being a multiple or one of their family members. In total, 3,264 blood samples from 3,221 NTR participants were assessed for genome-wide DNAm, including monozygotic and dizygotic twins, parents of twins, siblings of twins and spouses of twins. Informed consent was obtained from all participants. The study was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance- FWA00017598; IRB/institute codes, NTR 03-180).

DNAm assays

Blood sampling procedures have been described in detail previously[71]. DNAm was assessed with the Infinium HumanMethylation450 BeadChip Kit (Illumina, San Diego, CA, USA Illumina, San Diego, CA, USA) by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (http://www.glimdna.org/) as part of the Biobank-based Integrative Omics Study (BIOS) consortium[52]. DNAm measurements have been described previously[52,72]. Genomic DNA (500ng) from whole blood was bisulfite treated using the Zymo EZ DNA Methylation kit (Zymo Research Corp, Irvine, CA, USA), and 4 µl of bisulfite-converted DNA was measured on the Illumina 450k array following the manufacturer's protocol. QC of the raw DNAm data and normalization was performed according to the godmc pipeline.

Genotyping assays

Genotyping was done on multiple platforms, with a number of overlapping participants. The following platforms were chronologically used: Affymetrix-Perlegen, Illumina 660, Illumina Omni Express 1 M and Affymetrix 6.0. Genotype calls were made with the platform specific software (Birdseed, APT-Genotyper, Beadstudio) following manufacturers' protocols. For the Affymetrix-Perlegen and Illumina 660 platforms, the SNPs were lifted over to build 37 (HG19) of the Human reference genome. Per platform, a sample was removed if the call rate for this person was <90%, the Plink 1.07 inbreeding value F was <-0.075 or >0.075, the gender of the person did not match the DNA of the person, the IBD status did not match the expected familial relations, or the sample had more than mean+5sd Mendelian errors. For the Affymetrix 6.0 platform also all samples with a CQC value < 0.40 were removed. Afterwards, in case a subject, was genotyped on multiple platforms, only the platform with the highest number of SNPs was selected when concordance between platforms was over 97%. Allele - and strand alignment of SNPs was done against the Dutch GONL reference panel for each platform[73]. SNPs were removed in each platform when MAF < 0.005, HWE p < 1e-12 and the call rate of the SNP was < 95%[74]. Then SNPs were only selected if the allele frequency of the SNP deviated <0.10 as compared to the GONL data. Subsequently, the individual platform data was merged into a single dataset. In this single dataset, the sample IBD, on a common backbone of ~70K SNPs, was re-compared with their expected familial relations and samples were removed if they did not match. The single merged dataset was imputed with mach-admix, using GONL as a reference panel, for only the SNPs that survived QC and were present on at least one platform, forcing missing genotype imputation for all SNPs. Best guess genotypes were generated from these data and from these cross-platform imputed SNPs, the following SNPs were selected: SNPs with a R2>0.90, with HWE p > 0.00001, with a Mendelian error rate < 2% and if the association of one platform=case vs. the other platforms=controls p-value > 0.00001 (of course applied for each platform). This left 1.2M SNPs. These SNPs were then re-aligned against the 1000 Genomes Phase 3v5 reference and then imputed to that reference on the Michigan imputation Server[26]. From the resulting VCF files, best guess genotypes were calculated.

Data availability

DNAm data are available upon request in the European Genome-phenome Archive (EGA), under the accession code EGAD00010000887.

Acknowledgements

We would like to thank the twins and their family members for their participation.

Funding

Funding was obtained from the Netherlands Organization for Scientific Research (NWO) and The Netherlands Organisation for Health Research and Development (ZonMW) grants 904-61-090, 985-10-002, 904-61-193,480-04-004, 400-05-717, Addiction-31160008, 016-115-035, Middelgroot-911-09-032, NWO-Groot 480-15-001/674, Center for Medical Systems Biology (CSMB, NWO Genomics), NBIC/BioAssist/RK(2008.024), Biobanking and Biomolecular Resources Research Infrastructure (BBMRI –NL, 184.021.007 and 184.033.111); Spinozapremie (NWO- 56-464-14192), KNAW Academy Professor Award (PAH/6635) and University Research Fellow grant (URF) to DIB; Amsterdam Public Health research institute (former EMGO+) , Neuroscience Amsterdam research institute (former NCA) ; the European Science Foundation (ESF, EU/QLRT-2001-01254), the European Community's Seventh Framework Program (FP7- HEALTH-F4-2007-2013, grant 01413: ENGAGE); the European Research Council (ERC Starting 284167, ERC Consolidator 771057, ERC Advanced 230374), Rutgers University Cell and DNA Repository (NIMH U24 MH068457-06), the National Institutes of Health (NIH, R01D0042157-01A1, MH081802, DA018673, R01 DK092127-04, and Grand Opportunity grants 1RC2 MH089951); the Avera Institute for Human Genetics, Sioux Falls, South Dakota (USA). Part of the genotyping and analyses were funded by the Genetic Association Information Network (GAIN) of the Foundation for the National Institutes of Health. Computing was supported by NWO through grant 2018/EW/00408559, BiG Grid, the Dutch e-Science Grid and SURFSARA.

Prevention and Incidence of Asthma and Mite Allergy (PIAMA)

The PIAMA (The prevention and incidence of asthma and mite allergy) study is a birth cohort study of children born between 1996-1997. Details of the study design have been published previously[75]. In brief, pregnant women were recruited during their first trimester from the general population in 1996-1997 through antenatal clinics in the north, west and center of the Netherlands Non-allergic pregnant women were invited to participate in a ‘‘natural history’’ study arm. Pregnant women identified as allergic through the screening questionnaire were allocated primarily to an intervention arm with a random subset allocated to the natural history arm. The intervention involved the use of mite-impermeable mattress and pillow covers.

The study started with 3,963 newborns. Parents completed questionnaires on demographic factors, risk factors for asthma and respiratory symptoms at the child’s age of 3 months, annually from 1 to 8 years of age, and at 11, 14, and 16 years of age. Clinical examinations were performed in subgroups at ages 4, 8, 12 and 16 years. Written informed consent was obtained from the parent or legal guardians of the participants and from the participants themselves. The Medical Ethical Committees of the participating institutes approved the study. The PIAMA study Age 16 subjects were granted by Utrecht, METC (Medisch Ethische Toetsingscommissie) protocol number 12-019/K, May 25th 2012, Amendment 1, July 12th 2012, Amendment 2, September 20th 2012.

DNAm assays

In the PIAMA study, DNA from peripheral blood samples from 16 years PIAMA Children was extracted using the QIAamp blood kit (Qiagen or equivalent protocols), followed by precipitation-based concentration using GlycoBlue (Ambion). DNA concentration was determined by Nanodrop measurement and Picogreen quantification. 500 ng of DNA was bisulfite-converted using the EZ 96-DNA methylation kit (Zymo Research), following the manufacturer's standard protocol. After verification of the bisulfite conversion step using Sanger Sequencing, DNA concentration was normalized and the samples were randomized to avoid batch effects. DNAm was measured using the Illumina Infinium HumanMethylation450 beadchip. Each chip included one control DNA sample for QC purposes. Sample QC and normalisation was performed using *meffil*[9] in R.

Genotyping assays

DNA was collected from 2,162 children. Genome-wide genotyping was performed in three phases. The first phase was performed within the framework of the GABRIEL Consortium using an Illumina Human 610K quad array[17]. Genotypes were available from 172 children with asthma and from 187 controls after QC. A second group of 268 children who were more extensively examined during follow up was genotyped with an Illumina HumanOmniExpress array. A final group of 1,377 children was genotyped with the Illumina Human Omni Express Exome Array. SNPs were harmonized by base pair position annotated to genome build 37, name and annotation of strand for each platform. Discordant or duplicate SNPs or SNPs that showed large differences in allele frequencies between the 3 arrays (> 15 %) were removed. After QC, a total of 1,968 individuals remained and imputation was performed on the overlapping SNPs using the Michigan Imputation Server[26] with reference panel 1000G Phases 3[13]. We removed SNPs with imputation quality score <0.8, and also those SNPs with MAF <0.01. The final imputed dataset contained 6,687,384 SNPs.

Data availability

PIAMA data are available upon request. Requests can be submitted to the PIAMA Principal Investigators (https://piama.iras.uu.nl/english/).

Acknowledgements

The authors thank all the children and their parents for their cooperation. The authors also thank all the field workers and laboratory personnel involved for their efforts, and Marjan Tewis for data management.

Funding

The PIAMA study is supported by The Netherlands Organization for Health Research and Development; The Netherlands Organization for Scientific Research; The Lung Foundation of the Netherlands (grant number AF 45.1.14.001 supported the DNAm assays); BBMRI-NL, The Netherlands Ministry of Spatial Planning, Housing, and the Environment; and The Netherlands Ministry of Health, Welfare, and Sport.

PRECISESADS

The PRECISESADS project (www.precisesads.eu) aims to gather patients with systemic autoimmune diseases (SADs) into clusters taking in account OMICS information collected from peripheral blood cells, sera and urines of 2.500 people that will be further characterized with immunological and clinical data [76,77]. The recruitment of patients and samples for the PRECISESADS cross-sectional cohort has been done in different phases. The full set will consist of 2000 patients with systemic autoimmune diseases (SADs) and 666 healthy controls. The first phase is intended to recruit in deep and detailed molecular information on 288 individuals (48 patients for each disease and 48 controls). In the second phase, the remaining 2378 recruited individuals will be studied with validated and economically accessible assays. Written informed consent has been obtained for all PRECISESADS participants. Ethical approval for the study was obtained from the Local Research Ethics Committees.

DNAm assays

Infinium HumanMethylation450 BeadChips (Illumina, Inc., San Diego, CA, USA) were used to analyze DNAm of 528 PRECISESADS participants divided in three batches. DNA samples were bisulfite-converted using the EZ DNA methylation kit (Zymo Research, Orange, CA, USA). After bisulfite treatment, the remaining assay steps were performed following the specifications and using the reagents supplied and recommended by the manufacturer. The array was hybridized using a temperature gradient program, and arrays were imaged using a BeadArray Reader (Illumina Inc., San Diego, CA, USA).
Sample QC and functional normalization[10] was completed using with *meffil*[9] in R. Briefly, during QC steps we removed subjects based on genotype concordance, sex mismatches, outliers for methylated vs unmethylated signals, deviation from mean values at control probes, and high proportion of undetected probes. In total, we removed 12 subjects in our QC.

Genotyping assays

PRECISESADS participants were genotyped using the Illumina Human Core 24 v1a genome-wide SNP genotyping platform (Illumina Inc., San Diego, CA, USA) by the GENYO Institute (GENYO; Granada, Spain). Individuals were excluded on the basis of incorrect gender assignment, high missingness (>10 %) and non-European ancestry (using clinical information). Genotypes were filtered before imputation due to high missingness (>2%) HWE $p<0.001$, MAF < 1% and AT/CG changes with MAF > 40%. Following QC, the final directly genotyped dataset contained 212,868 SNP loci of 779 individuals. PLINK (v1.07)[11] was used to carry out QC measures on an initial set of 804 subjects and 306,670 directly genotyped SNPs. Genotypes were phased and imputed in one step using IMPUTE2 (version 2.3.2)[14] against the 1000 Genomes reference panel (phase 3)[13]. Genotypes were filtered after imputation to have HWE p > 1e-7, MAF >1 % and imputation info score >0.9 and resulted in 2,259,093 imputed genotypes.

Data availability

Data will be made available at the end of the project.

Acknowledgements

We are in debt to the PRECISESADS Clinical Consortium for their contribution in patient recruitment.

PRECISESADS Clinical Consortium members

Lorenzo Berettai[1], Barbara Vigone[1], Jacques-Olivier Pers[2], Alain Saraux[2], Valérie Devauchelle-Pensec[2], Divi Cornec[2], Sandrine Jousse-Joulin[2], Bernard Lauwerys[3], Julie Ducreux[3], Anne-Lise Maudoux[3], Carlos Vasconcelos[4], Ana Tavares[4], Esmeralda Neves[4], Raquel Faria[4], Mariana Brandão[4], Ana Campar[4], António Marinho[4], Fátima Farinha[4], Isabel Almeida[4], Miguel Angel Gonzalez-Gay Mantecón[5], Ricardo Blanco Alonso[5], Alfonso Corrales Martínez[5], Ricard Cervera[6], Ignasi Rodríguez-Pintó[6], Gerard Espinosa[6], Rik Lories[7], Ellen De Langhe[7], Nicolas Hunzelmann[8] Doreen Belz[8], Torsten Witte[9], Niklas Baerlecken[9], Georg Stummvoll[10], Michael Zauner[10], Michaela Lehner[10], Eduardo Collantes[11], Rafaela Ortega-Castro[11], Mª Angeles Aguirre-Zamorano[11], Alejandro Escudero-Contreras[11], Mª Carmen Castro-Villegas[11], Norberto Ortego[12], María Concepción, Fernández Roldán[12], Enrique Raya[13], Inmaculada Jiménez Moleón[13], Enrique de Ramon[14], Isabel Díaz Quintero[14], Pier Luigi Meroni[15], Maria Gerosa[15], Tommaso Schioppo[15], Carolina Artusi[15], Carlo Chizzolini[16], Aleksandra Zuber[16], Donatienne Wynar[16], Laszló Kovács[17], Attila Balog[17], Magdolna Deák[17], Márta Bocskai[17], Sonja Dulic[17], Gabriella Kádár[17], Falk Hiepe[18], Velia Gerl[18], Silvia Thiel[18], Manuel Rodriguez Maresca[19], Antonio López-Berrio[19], Rocío Aguilar-Quesada[19], Héctor Navarro-Linares[19]

PRECISESADS Clinical Consortium affiliations

[1] Referral Center for Systemic Autoimmune Diseases, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico di Milano, Italy.
[2] Centre Hospitalier Universitaire de Brest, Hospital de la Cavale Blanche, Brest, France.
[3] Pôle de pathologies rhumatismales systémiques et inflammatoires, Institut de Recherche Expérimentale et Clinique, Université catholique de Louvain, Brussels, Belgium.
[4] Centro Hospitalar do Porto, Portugal.
[5] Servicio Cantabro de Salud, Hospital Universitario Marqués de Valdecilla, Santander, Spain.
[6] Hospital Clinic I Provicia, Institut d'Investigacions Biomèdiques August Pi i Sunyer, Barcelona, Spain.
[7] Katholieke Universiteit Leuven, Belgium.
[8] Klinikum der Universitaet zu Koeln, Cologne, Germany.
[9] Medizinische Hochschule Hannover, Germany.
[10] Medical University Vienna, Vienna, Austria.
[11] Servicio Andaluz de Salud, Hospital Universitario Reina Sofía Córdoba, Spain.
[12] Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Universitario San Cecilio), Spain.
[13] Servicio Andaluz de Salud, Complejo hospitalario Universitario de Granada (Hospital Virgen de las Nieves), Spain.
[14] Servicio Andaluz de Salud, Hospital Regional Universitario de Málaga, Spain
[15] Università degli studi di Milano, Milan, Italy.
[16] Hospitaux Universitaires de Genève, Switzerland.
[17] University of Szeged, Szeged, Hungary.

[18] Charite, Berlin, Germany.
[19] Andalusian Public Health System Biobank, Granada, Spain

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement n° [115565], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

PREDO study

Data were from the Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) Study, which is a longitudinal multicenter pregnancy cohort study of Finnish women and their singleton children born alive between 2006-2010. We recruited 1079 pregnant women, of whom 969 had one or more and 110 had none of the known risk factors for preeclampsia and intrauterine growth restriction. The recruitment took place in arrival order when these women attended the first ultrasound screening at 12+0-13+6 weeks+days of gestation in one of the ten hospital maternity clinics participating in the study. The cohort profile contains details of the study design and inclusion criteria[78].

DNAm assays

Following DNA extraction, genomic DNA was bisulfite modified using an EZ DNA methylation kit (Zymo Research, Orange, CA, USA). The protocol was as described by the manufacturer. Genome-wide DNAm was measured using the Illumina HumanMethylation450 BeadChip and HumanInfinium Chip(Illumina, San Diego, CA, USA) following the manufacturer's protocol. Sample QC and normalisation was performed using *meffil*[9] in R.

Genotyping assays

Genotyping was performed on Illumina Human Omni Express Exome Arrays containing 964,193 SNPs. Only markers with a call rate of at least 98%, a MAF of 1% and a p-value for deviation from HWE p > 1e-6 were kept in the analysis. After QC, 587,290 SNPs were available. In total, 996 cord blood samples were genotyped. IDs with a call rate below 98% (n=11) were removed. Any ID-pair with IBD estimates > 0.125 was checked for relatedness. For most pairs, high IBD estimates could be explained due to African origin of these IDs. As we correct for admixture in our analyses, these IDs were kept except for one pair which could not be resolved. From this pair we excluded one ID from further analysis. Individuals showing discrepancies between phenotypic and genotypic sex (n=1) were removed. We also checked for heterozygosity outliers but found none. 983 IDs were available in the final dataset.

Data availability

Collaboration in data analysis is possible through specific research proposals sent to the PREDO Study Board [predo.study@helsinki.fi] or primary investigators Katri Räikkönen [katri.raikkonen@helsinki.fi] or Hannele Laivuori [hannele.laivuori@helsinki.fi].

Acknowledgements

The PREDO study would not have been possible without the dedicated contribution of the PREDO study group members: E Hamäläinen, E Kajantie, H Laivuori, PM Villa, A-K Pesonen, A Aitokallio-Tallberg, A-M Henry, VK Hiilesmaa, T Karipohja, R Meri, S Sainio, T Saisto, S Suomalainen-Konig, V-M Ulander, T Vaitilo (Department of Obstetrics and Gynaecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland), L Keski-Nisula, Maija-Riitta Orden (Kuopio University Hospital, Kuopio Finland), E Koistinen, T Walle, R Solja (Northern Karelia Central Hospital, Joensuu, Finland), M Kurkinen (Päijät-Häme Central Hospital, Lahti, Finland), P.Taipale. P Staven (Iisalmi Hospital, Iisalmi, Finland), J Uotila (Tampere University Hospital, Tampere, Finland). We thank all the PREDO children and their parents for their enthusiastic participation. We also thank all the research nurses, research assistants, and laboratory personnel involved in the Predo study.

Funding

The PREDO Study has been funded by the Academy of Finland, EraNet Neuron, EVO (a special state subsidy for health science research), University of Helsinki Research Funds, the Signe and Ane Gyllenberg foundation, the Emil Aaltonen Foundation, the Finnish Medical Foundation, the Jane and Aatos Erkko Foundation, the Novo Nordisk Foundation, the Päivikki and Sakari Sohlberg Foundation, the Sigrid Juselius Foundation granted to members of the Predo study board. DNAm assays were funded by the Academy of Finland.

Project MinE

These data are part of ProjectMinE as previously described[79]. In brief, project MinE is a collaboration of (inter)national groups with the aim to collect 22,500 DNA profiles to investigate rare and common (epi)genetic variation contributing to the development of Amyotrophic Lateral Sclerosis (ALS). The participants of this study consisted of a subset of 499 individuals of Dutch nationality and were ascertained when lacking any neurological phenotype. All participants gave written informed consent and the institutional review board of the University Medical Center Utrecht approved this study.

DNAm assays

DNAm status of the participants was extracted from whole blood samples using Illumina Infinium HumanMethylation450 BeadChip array following manufacturers protocol. Data was analyzed and normalized using *meffil*[9] in R version 3.3.3. Briefly, DNAm quality was checked by: the median intensity methylated vs unmethylated signal for all control probes (N=6), bisulfate 1 probes (N=0), bisulfate II probes (N=2), dyebias (N=0), detection pvalue (N=0), low bead numbers (N=0). Furthermore, data was checked for sex outliers and mismatches (N=7) and genotype mismatches (N=2) based on 65 SNP probes. After QC, the 623 remaining samples were normalized using 15 control probe PCs derived from the scree plot.

Genotyping assays

Genotyping was performed as previously described[5]. It should be noted that the 1000 Genomes Project v3 reference panel[13], and not the custom reference panel as indicated by

[79], was used for imputation. In total, 499 participants had matching genotype, DNAm and covariate data which could be used for further analyses.

Data availability

Data is available upon request at project MinE (https://www.projectmine.com).

Acknowledgements

The authors want to thank the study participants that contributed whole blood for this study. We further like to acknowledge Wouter van Rheenen for creating the imputed genotype data.

Funding

This study was supported by the ALS Foundation Netherlands.

The Western Australian Pregnancy Cohort (Raine) Study

The Western Australian Pregnancy Cohort (Raine) Study is a prospectively recruited longitudinal pregnancy cohort that recruited 2900 mothers between 1989 and 1991 from Western Australia's major perinatal centre, King Edward Memorial Hospital, and nearby private practices. Women who had sufficient English language skills, an expectation to deliver at King Edward Memorial Hospital, and an intention to reside in Western Australia to allow for future follow-up of their child were eligible for the study[80]. Informed consent was provided by all participants. Participant assent and parental consent was provided for minors.
The Raine Study has ethics approval from The University of Western Australia Human Research Ethics Committee.

DNAm assays

Processing of the Illumina Infinium HumanMethylation450 BeadChips was carried out by the Centre for Molecular Medicine and Therapeutics (CMMT) (http://www.cmmt.ubc.ca). Two packages were used to perform QC checks of the samples; *shinyMethyl*[81] and *MethylAid*[82]. Samples that were evident as outliers based on the output from *shinyMethyl* and *MethylAid* were removed. Gender discrepancy was inferred using the *RnBeads* package[83], and samples showing discrepancies were excluded. Samples that were run in duplicate or triplicate present on the Bead-Chip were used to assess genetic similarity between these individuals as a check for sample mix-ups. Samples showing signs of contamination based on heatmap produced by the *RnBeads* package[83] were excluded. Intentional SNP probes, sex chromosome probes and probes with a detection p-value greater than 0.05 in any sample were removed. Probes with low bead counts (bead count < 3 in more than 5% of samples) were also removed.
Post QC, the filtered and normalised data sets contain 1,255 samples and 462,927 probes.

Genotyping assays

The samples were genotyped using Illumina Human660W Quad Array at the Centre for Applied Genomics (Toronto, Ontario, Canada). PLINK (v1.07)[11] was used to carry out QC measures of the directly genotyped SNPs. Individual samples were tested for missingness (<0.05), heterozygosity (>-0.32) and genetic sex discrepancies. We also tested for individual IBS (pi > 0.1875), PCs were used to identify and remove ethnic outliers, and to adjust for population stratification in the downstream analyses. SNPs with MAF < 1%, genotype call rate <95% and/or a HWE with p<1e−6 were excluded, together with individuals with >5% missing genotypes. Samples that perform poorly in these QCs are excluded from further analysis. Phasing and Imputation was performed using Eagle v2.3[84] and Minimac3[26] respectively against the 1000 Genomes reference panel (phase 3, version 5)[13] using the Michigan Imputation Server[26]. The final imputed dataset contained SNPs. After removing duplicated SNPs, SNPs with MAF < 1%, and SNPs with poor imputation score (RSq < 0.80), a total number of 7,782,690 SNPs remained across the 22 autosomes for analytic purposes.

Data availability

Data is available upon request (https://ross.rainestudy.org.au).

Acknowledgements

The authors are grateful to the Raine Study participants and their families, and to the Raine Study research staff for cohort co-ordination and data collection. The authors gratefully acknowledge the following institutes for providing funding for Core Management of the Raine Study: The University of Western Australia (UWA), Curtin University of Technology, Murdoch University, The University of Notre Dame (Australia), the Raine Medical Research Foundation, the Telethon Kids Institute, the Women's and Infants' Research Foundation and Edith Cowan University. The authors gratefully acknowledge the assistance of the Western Australian DNA Bank (National Health and Medical Research Council of Australia National Enabling Facility).
This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and Government of Western Australia.

Funding

This study was supported by the National Health and Medical Research Council of Australia [grant numbers 403981, and 572613] and the Canadian Institutes of Health Research [grant number MOP-82893].

Rotterdam Study (RS)

The Rotterdam Study (RS) is a large prospective, population-based cohort study aimed at assessing the occurrence of and risk factors for chronic (cardiovascular, endocrine, hepatic, neurological, ophthalmic, psychiatric, dermatological, oncological, and respiratory) diseases in the elderly[85]. The study comprises 14,926 subjects in total, living in the well-defined Ommoord district in the city of Rotterdam in the Netherlands. In 1989, the first cohort, Rotterdam Study-I (RS-I) comprised of 7,983 subjects with age 55 years or above. In 2000, the second cohort, Rotterdam Study-II (RS-II) was included with 3,011 subjects who had

reached an age of 55 or over in 2000. In 2006, the third cohort, Rotterdam Study-III (RS-III) was further included with 3,932 subjects with age 45 years and above. Each participant gave an informed consent and the study was approved by the medical ethics committee of the Erasmus University Medical Center, Rotterdam, the Netherlands.

DNAm assays

At the Genetic Laboratory, Human Genomics Facility (HuGe-F), Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, the Netherlands, the DNAm dataset was generated for a subset of 747 individuals of RS-III at baseline. The second DNAm dataset was generated in another subset of 864 individuals comprising of individuals at their fifth, third and second visit of RS-I, RS-II and RS-III respectively, between 2009-2013. Genomic DNA was extracted from whole peripheral blood by standardized salting out methods. This was followed by a bisulfite conversion using the Zymo EZ-96 DNA-methylation kit (Zymo Research, Irvine, CA, USA). The genome for each sample was then amplified, fragmented and hybridized to the Infinium Illumina Human Methylation 450k arrays according to the manufacturer's protocol.

Genotyping assays

SNP data was obtained from the Illumina HapMap 550k platform in 11,496 samples of European ancestry in the Rotterdam Study. Genotype calling was done using GenomeStudio. Samples were excluded based on sample call rate ≤97.8, heterozygosity > median + 3*IQR, gender mismatches, duplicates, PCA outliers using a PCA projection of the study samples onto 1KG reference samples, ethnic outliers of 4SD CEU HapMap and Cryptic Relatedness of PI HAT>0.3. SNPs were excluded based on ≤ 1 % MAF, HWE P value < 1e-6 and call rate ≤ 98%. Imputations were performed using the MACH/Minimac software[18,26].

Data availability

Requests for the data accession may be sent to: Frank van Rooij (f.vanrooij@erasmusmc.nl).

Acknowledgements

We thank Mr. Michael Verbiest, Ms. Mila Jhamai, Ms. Sarah Higgins and Mr. Marijn Verkerk for their help in creating the DNAm database. We thank Pascal Arp, Mila Jhamai, Marijn Verkerk, Lizbeth Herrera and Marjolein Peters, MSc, and Carolina Medina-Gomez, MSc, for their help in creating the GWAS database, and Karol Estrada, PhD, Yurii Aulchenko, PhD, and Carolina Medina-Gomez, MSc, for the creation and analysis of imputed data. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

Funding

The EWAS data was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by the Netherlands Organization for Scientific Research (NWO;

project number 184021007) and made available as a Rainbow Project (RP3; BIOS) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL).

The GWAS datasets are supported by the Netherlands Organisation of Scientific Research NWO Investments (nr. 175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project nr. 050-060-810.

The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam.

This work was done within the framework of the Biobank-Based Integrative Omics Studies (BIOS) Consortium funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007).

The Southall And Brent REvisited Study (SABRE)

The Southall And Brent REvisited Study (SABRE) is a population based cohort including 1710 first generation South Asian migrants, 801 first generation African Caribbean migrants, and 2346 people of European origin aged 40 to 69 living in West London, UK[86]. Baseline investigations were performed between 1988 and 1991. Peripheral blood samples were collected at baseline visits from participants recruited in the Southall district.
The study was approved by St Mary's Hospital Research Ethics Committee (07/H0712/109) and all participants provided written informed consent. The study adheres to the principles of the Declaration of Helsinki and Title 45, US Code of Federal Regulations, Part 46, Protection of Human Subjects, Revised November 13, 2001, effective December 13, 2001.

DNAm assays

Following DNA extraction, genomic DNA (500 ng) was bisulfite modified using an EZ DNA methylation kit (Zymo Research, Orange, CA, USA). The protocol was as described by the manufacturer, utilising the alternative incubation conditions recommended when using Illumina Infinium Methylation Arrays. Genome-wide DNAm was measured using the Illumina HumanMethylation450 BeadChip (Illumina, San Diego, CA, USA) following the manufacturer's protocol with no modifications. The arrays were scanned using an Illumina iScan with software version 3.3.28.
Sample QC and normalisation was performed using *meffil*[9] in R version 3.1.1. Briefly, DNAm quality was evaluated by: sex detection outliers (N=13), the median intensity methylated vs unmethylated signal (N=10), dyebias (N=21), DNAm detection pvalue (N=588), and low bead numbers (N=167). Finally, 1993 samples passed QC. SABRE was normalized using 15 control probe PCs derived from the technical probes informed by *meffil* scree plots.
After filtering for European ancestry, those with DNAm measured at baseline, and those with genetic data available there were 731 individuals and 484,781 DNAm probes available for analysis.

Genotyping assays

2980 SABRE participants from the Southall centres were genotyped using a new UCL druggable target array. The array has a GWAS backbone comprising the Illumina Human Core BeadChip (~240k genome wide markers) and an additional custom set of 200k markers on genes encoding proteins involved in drug handling, drug action, and druggable targets. This was developed in collaboration with LSHTM and EBI.

Individuals were excluded due to incorrect sex assignment, high missingness (>5%), abnormal heterozygosity (het > mean (het) + 3*sd (het) or het < mean (het) - 3*sd (het)), cryptic relatedness (pihat>=0.9999 identified 20 cases of sample mislabelling (excluded), Pihat = 0.5 → 1st degree relative. Pihat = 0.25 → 2nd degree relative. – these were known relatives and remain in the dataset), and non-European ancestry (detected via PCA). Phasing was performed using Eagle and ethnicity-specific imputation was performed using Minimac3[18] against the HRC reference panel (http://www.haplotype-reference-consortium.org/)[87]. Genotypes were filtered to have a MAF >0.005, imputation info score >0.3 and HWE p<0.00001. After data cleaning, QC, filtering for European ancestry, and sample matching to DNAm data availability, 731 individuals and 6,912,559 SNPs remained. A further 9 samples were removed by the GoDMC pipeline.

Data availability

Data are available by request from SABRE (https://www.sabrestudy.org).

Acknowledgements

450K DNAm array data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Funding

The SABRE study was funded at baseline by the Medical Research Council, Diabetes UK, and the British Heart Foundation. DNAm analysis in the SABRE cohort was supported by a Wellcome Trust Enhancement grant 082464/Z/07/C. Genotyping analysis in the SABRE cohort was supported by the British Heart Foundation (CS/13/1/30327)

Schizophrenia Phase 1- University College London (SCZ1)

The University College London case-control sample has been described elsewhere[88] but briefly comprises of unrelated ancestrally matched schizophrenia cases and controls from the United Kingdom. Case participants were recruited from UK NHS mental health services with a clinical ICD-10 diagnosis of schizophrenia. All case participants were interviewed with the Schedule for Affective Disorders and Schizophrenia-Lifetime Version (SADS-L) to confirm Research Diagnostic Criteria (RDC) diagnosis. A control sample screened for an absence of mental health problems was recruited. Each control subject was interviewed to confirm that they did not have a personal history of an RDC defined mental disorder or a family history of schizophrenia, bipolar disorder, or alcohol dependence. All subjects signed an approved consent form after reading an information sheet. UK National Health Service multicentre and local research ethics approval was obtained.

DNAm assays

Illumina 450K BEadchip arrays were processed at the University of Exeter Medical School laboratories by the Complex Disease Epigenetics Group as previously described[89]. 500ng of DNA from each sample was treated with sodium bisulfite in duplicate, using the EZ-96 DNA Methylation kit (Zymo Research, CA, USA). DNAm was quantified using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc, CA, USA) run on an Illumina iScan System (Illumina, CA, USA) using the manufacturers' standard protocol. Samples were randomly assigned to chips and plates to ensure equal distribution of cases and controls across arrays and minimise batch effects. In addition, a fully methylated control (CpG Methylated HeLa Genomic DNA; New England BioLabs, MA, USA) was included in a random position on each plate.

Genotyping assays

Genotyping was performed using the Affymetrix Mapping 500K Array and the Genomewide Human SNP Array 5.0 or 6.0 (Affymetrix, CA, USA). Genotypes were called from raw intensity data using the Birdseed component of the Birdsuite algorithm. Samples were genotyped by the Genetic Analysis Platform at The Broad Institute of Harvard and MIT according to standard protocols. Prior to imputation PLINK[11] was used to remove samples with >5% missing data. We also excluded SNPs characterized by >5% missing values, a HWE $p < 0.001$ and a MAF<5%. Imputation was performed using ChunkChromosome (http://genome.sph.umich.edu/wiki/ChunkChromosome) and Minimac2[18] with the 1000 Genomes reference panel of European samples (phase 1, version 3)[13].

Data availability

DNAm data are available through GEO under accession number GSE80417.

Funding

This work was primarily supported by a grant from the UK Medical Research Council (MRC; MR/K013807/1) to J.M. and the US National Institutes of Health (NIH) (R01 AG036039) to J.M.

Schizophrenia Phase 2- Aberdeen (SCZ2)

The Aberdeen case-control sample has been described elsewhere[90] but briefly contains schizophrenia cases and controls who have self-identified as born in the British Isles (95% in Scotland). All cases met the Diagnostic and Statistical Manual for Mental Disorders-IV edition (DSM-IV) and International Classification of Diseases 10th edition (ICD-10) 2 criteria for schizophrenia. Diagnosis was made by Operational Criteria Checklist (OPCRIT). All case participants were outpatients or stable in-patients. Detailed medical and psychiatric histories were collected. A clinical interview using the Structured Clinical Interview for DSM-IV (SCID) was also performed on schizophrenia cases. Controls were volunteers recruited through general practices in Scotland. Practice lists were screened for potentially suitable volunteers by age and sex and by exclusion of subjects with major mental illness or use of neuroleptic medication. Volunteers who replied to a written invitation were interviewed using a short questionnaire to exclude major mental illness in individual themselves and first degree

relatives. All cases and controls gave informed consent. The study was approved by both local and multiregional academic ethical committees.

DNAm assays

Illumina 450K Beadchip arrays were processed at the University of Exeter Medical School laboratories by the Complex Disease Epigenetics Group as previously described[89] 500ng of DNA from each sample was treated with sodium bisulfite in duplicate, using the EZ-96 DNA Methylation kit (Zymo Research, CA, USA). DNAm was quantified using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc, CA, USA) run on an Illumina iScan System (Illumina, CA, USA) using the manufacturers' standard protocol. Samples were randomly assigned to chips and plates to ensure equal distribution of cases and controls across arrays and minimise batch effects. In addition, a fully methylated control (CpG Methylated HeLa Genomic DNA; New England BioLabs, MA, USA) was included in a random position on each plate.

Genotyping assays

Genotyping was performed using the Affymetrix Mapping 500K Array and the Genomewide Human SNP Array 5.0 or 6.0 (Affymetrix, CA, USA). Genotypes were called from raw intensity data using the Birdseed component of the Birdsuite algorithm. Samples were genotyped by the Genetic Analysis Platform at The Broad Institute of Harvard and MIT according to standard protocols. Prior to imputation PLINK[11] was used to remove samples with >5% missing data. We also excluded SNPs characterized by >5% missing values, a HWE $p < 0.001$ and a MAF of <5%. Imputation was performed using ChunkChromosome (http://genome.sph.umich.edu/wiki/ChunkChromosome) and Minimac2[18] with the 1000 Genomes reference panel of European samples (phase 1, version 3)[13].

Data availability

DNAm data are available through GEO under accession number GSE84727.

Funding

This work was primarily supported by a grant from the UK Medical Research Council (MRC; MR/K013807/1) to J.M. and the US National Institutes of Health (NIH) (R01 AG036039) to J.M.

Saguenay Youth Study (SYS)

The Saguenay Youth Study (SYS) cohort includes 1029 adolescents and 962 parents. The cohort was recruited via adolescents attending high schools in the Saguenay-Lac-Saint-Jean region of Quebec, Canada[91,92]. Written informed consents of adults and assents of adolescents (and consents of their parents) have been obtained from all SYS participants. The regional ethics committees approved the study protocols.

DNAm assays

Epityping was conducted on DNA extracted from peripheral blood cells using the Infinium HumanMethylation450K BeadChip (Illumina) at the SNP&SEQ Technology Platform, Uppsala University (Uppsala, Sweden). DNAm β values were normalized using Genome Studio. QC was performed by excluding sites with detection P < 0.01. All samples had >98% sites with detection P < 0.01.

Genotyping assays

The SYS adolescents and parents were genotyped in two waves. First, 592 adolescents were genotyped with the Illumina Human610-Quad BeadChip (Illumina; n = 582,892 SNPs) at the Centre National de Génotypage (Paris, France). Second, the remaining 427 adolescents and all parents were genotyped with the HumanOmniExpress BeadChip (Illumina; n = 729,295 SNPs) at the Genome Analysis Centre of Helmholtz Zentrum München (Munich, Germany). In both genotyping waves, SNPs with call rate <95% and MAF <0.01 and SNPs that were not in HWE ($p < 1e{-}6$) were excluded. After this QC, 542,345 SNPs on the first chip and 644,283 SNPs on the second chip were available for analysis.

Genotype imputation was used to equate the set of SNPs genotyped on each platform and to increase the SNP density. Haplotype phasing was performed with SHAPEIT[12] using an overlapping subset of 313,653 post-QC SNPs that were present on both genotyping platforms and the 1000 Genomes SNPs in European reference panel (Phase 1, Release 3)[13]. Imputation was conducted on the phased data with IMPUTEv2[14]. Markers with low imputation quality (information score <0.5) or MAF<0.01 were removed. After this QC of imputation, a total of 7,746,837 typed and imputed SNPs were analyzed.

Data availability

Data are available upon request addressed to Dr Zdenka Pausova [zdenka.pausova@sickkids.ca] and Dr Tomas Paus [tpaus@research.baycrest.org]. Further details about the protocol can be found at [http://www.saguenay-youth-study.org/].

Acknowledgements

We thank all families who took part in the Saguenay Youth Study, and the whole SYS team.

Funding

The Saguenay Youth Study has been funded by the Canadian Institutes of Health Research (TP, ZP), Heart and Stroke Foundation of Canada (ZP) and the Canadian Foundation for Innovation (ZP). The McLaughlin Centre at the University of Toronto provided supplementary funds for the DNAm studies in the SYS. TP was supported by Tanenbaum Chair in Population Neuroscience (University of Toronto).

The UK Adult Twin (TwinsUK)

The UK Adult Twin Registry (TwinsUK) was established in 1992 to recruit MZ and DZ adult same-sex twins[93]. The majority of participants are healthy female Caucasians (age range

from 16 to 98 years old). The cohort consists of more than 13,000 twins from all regions across the United Kingdom, and many have multiple visits over the years, at which clinical measurements and biological samples have been collected. The study was approved by the St. Thomas' Hospital Research Ethics committee (EC04/015 and 07/H0802/84) and all subjects provided informed written consent.

DNAm assays

TwinsUK DNAm data and QC assessments have been previously described[94].

Genotyping assays

Genotyping and genotype data imputation have been previously described[95].

Data availability

DNAm data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE62992 and GSE121633. Access to additional individual-level genotype and phenotype data can be applied for through the TwinsUK data access committee http://twinsuk.ac.uk/resources-for-researchers/access-our-data/

Acknowledgements

We thank Dr Pei-Chien Tsai and Dr Leonie Roos for contributions to data QC, Dr Massimo M Mangino for genotype imputation, and Dr Kerrin S Small for discussion of data analysis approaches. We are grateful to all twins who participated in the study.

Funding

The study was supported by the UK Economic and Social Research Council (ES/N000404/1 to JTB) and JPI HDHL funded DIMENSION project, administered by BBSRC (BB/S020845/1 to JTB). The TwinsUK cohort is funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013) and also receives support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. SNP genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR.

UK Household Longitudinal Study (UKHLS - also known as Understanding Society)

The British Household Panel Survey (BHPS) began in 1991, and in 2010 was incorporated into the larger UK Household Longitudinal Study[96] (UKHLS; also known as Understanding Society (https://www.understandingsociety.ac.uk) which is a longitudinal panel survey of 40,000 UK households from England, Scotland, Wales and Northern Ireland. Since 1991 annual interviews have collected sociodemographic information, and in 2011-12, biomedical measures and blood samples for BHPS participants were collected at a nurse visit in the participant's home. Respondents were eligible to give a blood sample if they had taken part in the previous main interview in English, were aged 16+, lived in England, Wales or

Scotland, were not pregnant, and met other conditions detailed in the user guide (https://www.understandingsociety.ac.uk/sites/default/files/downloads/legacy/7251-UnderstandingSociety-Biomarker-UserGuide-2014-1.pdf). For each participant, non-fasting blood samples were collected through venepuncture; these were subsequently centrifuged to separate plasma and serum, aliquoted and frozen at −80 °C. DNA has been extracted and stored for genetic and epigenetic analyses.
Participants gave written consent for blood sampling. Ethical approval for the Understanding Society nurse visit was obtained from the National Research Ethics Service (Reference: 10/H0604/2).

DNAm assays

DNAm was profiled in DNA extracted from whole blood for 1,193 individuals aged from 28 to 98 who were eligible for and consented to both blood sampling and genetic analysis, had been present at all annual interviews between 1999 and 2011, and whose time between blood sample collection and processing did not exceed 3 days. Eligibility requirements for genetic analyses meant that the epigenetic sample was restricted to participants of white ethnicity. 500ng of DNA from each sample was treated with sodium bisulfite, using the EZ-96 DNA methylation-Gold kit (Zymo Research, CA, USA). DNAm was quantified using the Illumina Infinium HumanMethylationEPIC BeadChip (Illumina Inc, CA, USA) run on an Illumina iScan System (Illumina, CA, USA) using the manufacturers' standard protocol at the University of Exeter Medical School by the Complex Disease Epigenetics Group. Samples were randomly assigned to chips and plates to minimise batch effects. In addition, a fully methylated control (CpG Methylated HeLa Genomic DNA; New England BioLabs, MA, USA) was included in a random position on each plate to facilitate sample tracking, resolve experimental inconsistencies and confirm data quality. Raw signal intensities were imported from idats into the R statistical environment and converted into beta values using the *bigmelon* package[97]. Data was processed through a standard pipeline and included the following steps: outlier detection, confirmation of complete bisulfite conversion, estimating age from the data[3] and comparing with reported age visualisation of PCs. Data were normalized using the dasen function from the *wateRmelon* package[4]. Samples that were dramatically altered as a result of normalisation were excluded by assessing the difference between normalized and raw data and removing those with a root mean square and standard deviation > 0.05. were then filtered to exclude samples and then DNAm sites with > 1% of sites or samples with detection p value > 0.05, finally DNAm sites with a bead count < 3 were also excluded. The data were then re-normalised with the dasen function.

Genotyping assays

UKHLS samples were genotyped typed using the Illumina Infinium HumanCoreExome BeadChip Kit® (12v1-0)[98]. This array contains a set of >250,000 highly informative genome-wide tagging single nucleotide polymorphisms as well as a panel of functional (protein-altering) exonic markers, including a large proportion of low-frequency (MAF 1–5%) and rare (MAF <1%) variants. Genotype calling was performed with the gencall algorithm using GenomeStudio (Illumina Inc.). For QC individuals were excluded based on the following criteria: sample call rate <98%, autosomal heterozygosity outliers (>3SD), gender mismatches, duplicates as established by identity by descent (IBD) analysis (PI_HAT > 0.9). Individuals with non-European ancestry were also excluded. For this genomic kinship was

estimated between all pairs of individuals along with 1000 Genomes Project data. These were converted to distances and subjected to multidimensional scaling. Prior to variant QC, all 538,448 variants were mapped to the human reference genome build 37. Variants with HWE p-value < 1e−4, call rate below 98% or poor genotype clustering values (<0.4) were excluded, leaving 525,314 variants passing QC. After selecting only the samples with matched DNAm data, variants were refiltered prior to imputation. PLINK[11] was used to remove samples with >5% missing data. We also excluded SNPs characterized by >5% missing values, a HWE p-value < 0.001 and a MAF of <5%. These data were then imputed using the 1000 genomes phase 3 version 5 reference panel[13] using SHAPEIT[12] and minimac3[26].

Data availability

Individual level DNAm and genetic data are available on application through the European Genome-phenome Archive under accession EGAS00001001232 (https://www.ebi.ac.uk/ega/home). Specific details can be found here (https://www.understandingsociety.ac.uk/about/health/data). Phenotype linked to DNAm data are available through application to the METADAC (www.metadac.ac.uk).

Acknowledgements

Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. We acknowledge the Wellcome Trust Sanger Institute for generating the genotype data. Analysis was facilitated by access to the Genome high performance computing cluster at the University Of Essex School Of Biological Sciences.

Funding

Both genotyping and DNAm in UKHLS were funded through enhancements to the Economic and Social Research council (ESRC) grants ES/K005146/1 and ES/N00812X/1.

Replication Cohort

Generation Scotland: Scottish Family Health Study (GS:SFHS)

GS:SFHS is a population- and family-based cohort comprising ~24,000 individuals from the Scottish population recruited between 2006 and 2011. The cohort has been described in detail previously[99]. Blood samples for DNA extraction were obtained at the time of recruitment to GS:SFHS. Participants provided written consent. Research Tissue bank approval was provided by the East of Scotland Research Ethics Service (REC reference 15/ES/0040). Original sample collection ethical approval was provided by NHS committees on research ethics for the GS:SFHS (reference 05/s1401/89) and STRADL (reference 14/SS/0039) studies.

DNAm assays

Whole blood genomic DNA (500 ng) was treated with sodium bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, California), following the manufacturer's instructions. DNAm was assessed using the Infinium MethylationEPIC BeadChip (Illumina Inc., San Diego, California), according to the manufacturer's protocol. The arrays were scanned using a HiScan scanner (Illumina Inc., San Diego, California) and initial inspection of array quality was carried out using Genome Studio version 2011.1.

QC of the DNAm was carried out before normalisation. The R package *shinyMethyl* was used to perform preliminary QC[81]. The filtering process removed 81 samples, including 1) outliers based on overall array signal intensity and control probe performance, 2) samples showing mismatch between recorded gender and predicted gender based on X and Y chromosome DNAm, and 3) genetic ethnic outliers identified by a PCA for the GS:SFHS cohorts[100]. An additional QC was performed using the pfilter function in the R package *wateRmelon*[4] based on the following criteria: samples are removed if ≥ 1% sites have a detection P value of > 0.05. This removed 18 further samples, leaving 5,101 samples to be used in the normalization step. Before normalization, individual probe-sample pairs with a detection P value of > 0.05 were removed. Normalization was performed using function preprocessNoob in the R package *minfi*[39]. This produced data in the form of beta- and M-values.

In order to remove potential effects from technical factors, we used linear mixed modelling to pre-correct each QCed probe for fixed and random effects (fixed effects: top 50 PCs of control probe intensities (explaining 99% of variation in control probe intensities), appointment clinic centre, processing batch, year of the visit, and sentrix position (position of the sample in Illumina slide); random effects: the appointment date and the sentrix ID (Illumina slide). The model converged successfully for 712,595 sites, and the resultant residualised M-values were used as DNAm phenotypes in downstream analysis. For individual sites, outlier samples with residualized-M-values more than five interquartile ranges from the nearest quartile were removed. Cell counts were estimated for granulocytes, monocytes, B-lymphocytes, natural killer cells, CD4+ T-lymphocytes and CD8+ T-lymphocytes using the estimateCellCounts() function in R package *minfi*[39] and were corrected in later analysis stages (described in Methods). The residuals (following the fitting of the GKFSC model) were rank-based inverse-normal transformed.

Genotyping assays

Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland. Genotyping was carried out using Illumina HumanOmniExpressExome-8 v1.0 and HumanOmniExpressExome-8 v1.2 BeadChips with Infinium chemistry for both. Genotypes were processed using the IlluminaGenomeStudio Analysis software v2011.1 (Illumina, San Diego, CA) and called using Beadstudio-Gencall v3.0. The details of blood collection and DNA extraction are provided elsewhere[101]. QC removed individuals with <98% call rate, SNPs with <98% call rate, and SNPs with a HWE p <1e-6. After initial QC, 604,858 genotyped autosomal SNPs remained. The genotyped data were imputed utilising the Sanger Imputation Service using the HRC panel v1.1[87] as described previously[102]. The data (n=602,450 SNPs) were pre-phased using SHAPEIT v2.r873 + duohmm[12,103] and imputed with PBWT[104].

Data availability

Non-identifiable data from this study will be made available to researchers through GS:SFHS Access Committee.

Acknowledgements

We are grateful to the participants for their involvement in this study. KLE acknowledges the support of the Brain & Behavior Research Foundation through a NARSAD Independent Investigator Award, which led to the production of pilot data used in this work. KLE, RMW and AMM are members of The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology (CCACE), part of the cross council Lifelong Health and Wellbeing Initiative (G0700704/84698). Funding of CCACE from the BBSRC, EPSRC, ESRC and MRC is gratefully acknowledged. ADB acknowledges the support of the Wellcome Trust.

Funding

Genotyping was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award “STratifying Resilience and Depression Longitudinally” (STRADL) Reference 104036/Z/14/Z). The DNAm analysis was funded by a Wellcome Trust Strategic Award “Stratifying Resilience and Depression Longitudinally” (STRADL) (Reference 104036/Z/14/Z). AMM is additionally supported by the Sackler Foundation. 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]. ADB has been supported by a Wellcome Trust PhD Training Fellowship for Clinicians, the Edinburgh Clinical Academic Track (ECAT) programme (204979/Z/16/Z).

Isolated subsets

The Multi-Ethnic Study of Atherosclerosis (MESA) Epigenomics and Transcriptomics Study

The MESA was designed to investigate the prevalence, correlates and progression of subclinical cardiovascular disease in a population cohort of 6,814 participants. The MESA Epigenomics and Transcriptomics Study comprises 1,264 randomly selected MESA participants.[105] We downloaded the normalised DNAm data from GEO (GSE56046 and GSE56581). GSE56046 comprises DNAm profiles from CD14+ samples from 1,202 individuals ranging 44 - 83 years of age. Peripheral monocytes were isolated from blood with anti-CD14 coated magnetic beads. GSE56581 comprises DNAm profiles from CD4+ samples, collected from 214 individuals. Peripheral T cells were isolated from blood with anti-CD4 coated magnetic beads. The Illumina HumanMethylation450 BeadChip and HiScan reader were used to measure DNAm profiles. Data was normalised using quantile normalisation. We removed DNAm measurements that were 10 SD from the mean using 3 iterations.

Other tissues

GSE78743

Samples of 12 tissues from 16 obductions[8,106] were taken within 12 hours post-mortem (mean age, 62.8 years). Solid tissues were snap frozen for further processing and stored at -80°C. Whole blood from the thoracic cavity was stored in disodium salt dehydrate (EDTA) tubes (BD, United Kingdom). In concordance with the ethical guidelines in the Code for Proper Secondary Use of Human Tissue in the Netherlands (Dutch Federation of Medical Scientific Societies), these samples were anonymized, and raw data have been deposited in the National Center for Biotechnology Information's GEO (GSE78743).

DNAm assays

Genomic DNA from blood was isolated using the Qiagen Minikit (Qiagen, Hilden, Germany) according to manufacturer's protocol. Genomic DNA from tissues was isolated using phenol/chlorophorm extraction. Bisulfite-converted DNA was generated using the Zymo EZ DNA Methylation kit (Zymo, Irvine, CA, USA). DNA methylation was measured using the Illumina Infinium HumanMethylation450 BeadChip according to manufacturer's protocol. Initial QC was performed using the R package MethylAid[82]. Following QC, 152 samples were kept. Functional normalisation was performed using the package *meffil*[9]. We removed DNAm measurements that were 10 SD from the mean using 3 iterations.

Data availability

DNAm data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE78743.

Acknowledgements

450K DNAm array data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Brain samples

Human fetal brain tissue was acquired from the Human Developmental Biology Resource (HDBR) (http://www.hdbr.org) and MRC Brain Banks network (https://mrc.ukri.org/research/facilities-and-resources-for-researchers/brain-banks/how-to-access-brain-tissue/). Ethical approval for the HDBR was granted by the Royal Free Hospital research ethics committee under reference 08/H0712/34 and Human Tissue Authority (HTA) material storage licence 12220; ethical approval for MRC Brain Bank was granted under reference 08/MRE09/38. A detailed description of these samples can be found in[107,108]. Briefly 173 fetal brain samples (94 male, 79 female) ranging from 56 to 169 days post-conception were used for DNA methylation and SNP profiling. Brain tissue was obtained frozen and had not been dissected into regions. Half of the brain tissue from each individual fetus was homogenized for subsequent genomic DNA extraction. Postmortem brain specimens were collected postmortem following consent obtained with next of kin, dissected

by neuropathology technicians, snap-frozen and stored at -80°C. Genomic DNA was isolated from all brain samples using a standard phenol-chloroform extraction protocol. DNA was tested for degradation and purity using spectrophotometry and gel electrophoresis.

DNAm assays

500ng of DNA from each sample was treated with sodium bisulfite in duplicate, using the EZ-96 DNA Methylation kit (Zymo Research, CA, USA). DNA methylation was quantified using the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc, CA, USA) run on an Illumina iScan System (Illumina, CA, USA) using the manufacturers' standard protocol. Signal intensities for each probe were extracted using Illumina GenomeStudio software (Illumina, CA, USA) and imported into the R statistical programme using the methylumi and minfi packages[39]. Multi-dimensional scaling (MDS) plots of variable probes on the sex chromosomes were used to check that the predicted gender corresponded with the reported gender for each individual. Further data quality control and processing steps were conducted using the wateRmelon package[4] in R. The pfilter function was used to filter firstly samples with >1% probes with a detection P value > 0.05 were removed and probes with a detection P value > 0.05 in at least 1% samples or/and a beadcount < 3 in 5% of samples were removed across all samples to control for poor quality probes. The dasen function was used to normalize the data as previously described[4]. These data are publicly available through GEO and can be found under accession number GSE58885.
Before commencing QTL analyses, genotypes at the polymorphic SNP probes on the HumanMethylation 450K array were compared to calls from the HumanOmniExpress genotyping array to confirm sample identity.

Genotyping assays

200ng of genomic DNA from each sample was genotyped using the Illumina HumanOmniExpress BeadChip (Illumina Inc, CA, USA). Following scanning, Illumina GenomeStudio software was used for genotype calling and the data were exported as ped and map files. Prior to imputation PLINK[11] was used to remove samples with >5% missing data. We also excluded SNPs characterized by >1% missing values, a Hardy-Weinberg equilibrium P < 0.001 and a minor allele frequency of <5%. These were recoded as vcf files using PLINK1.94[109] and VCFtools[110] before uploading to the Michigan Imputation Server (https://imputationserver.sph.umich.edu/start.html#!pages/home) which uses SHAPEIT[12,111] to phase haplotypes, and Minimac3[18] with the most recent 1000 Genomes reference panel (phase 3, version 5). Imputed genotypes were then filtered and recoded with PLINK1.94[109] removing samples with >5% missing values, and SNPs with >2 alleles, those indicated as a fail in the FILTER columns using the flag '--vcf-filter', in addition to those characterized by >1% missing values, a Hardy-Weinberg equilibrium P < 0.001, a minor allele frequency of < 5%.

Funding

This work was supported by grants from the UK Medical Research Council (MRC) (grant numbers MR/K013807/1 and MR/L010674/1) to JM, and US National Institutes of Health (grant number AG036039) to JM. The human embryonic and fetal material was provided by the Joint MRC (grant number G0700089)/Wellcome Trust (grant number GR082557) Human Developmental Biology Resource.

TwinsUK adipose tissue

Adipose tissue samples from the TwinsUK cohort were collected as part of the MuTHER (Multiple Tissue Human Expression Resource) study[112]. For this study, the sample consisted of 596 healthy Caucasian adult female twins. Adipose tissue biopsies were obtained between August 2007 and May 2009. Ethical approval was granted by the National Research Ethics Service London-Westminster, the St Thomas' Hospital Research Ethics Committee (EC04/015 and 07/H0802/84). All research participants signed an informed consent.

DNAm assays

DNA methylation profiling was performed with the Illumina Infinium HumanMethylation450 BeadChip as described previously[113].

Genotype assays

Genotyping was done with a combination of Illumina arrays (HumanHap300, HumanHap610Q, 1M-Duo, and 1.2MDuo 1M)[95]. Imputation was performed using the IMPUTE software package (v2)[14] using as reference panel the 1000 Genomes haplotypes Phase I integrated variant set release (v3, September 2013).

Data availability

DNA methylation data are stored in the ArrayExpress repository (https://www.ebi.ac.uk/arrayexpress/) under the accession number E-MTAB-1866. Genotype data are available upon request through application to the TwinsUK data access committee (https://twinsuk.ac.uk/resources-for-researchers/access-our-data/).

Funding

TwinsUK was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013) and also receives support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. SNP genotyping was performed by the Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR. The MuTHER Study was funded by the Wellcome Trust (081917/Z/07/Z) and core funding for the Wellcome Trust Centre for Human Genetics (090532).

References

1. Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* (2018) doi:10.1038/s41431-018-0177-4.
2. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D.

Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, e90–e90 (2013).

3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
4. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
5. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).
6. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **42**, 111–127 (2013).
7. Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int. J. Epidemiol.* **42**, 97–110 (2013).
8. Relton, C. L. *et al.* Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int. J. Epidemiol.* **44**, 1181–1190 (2015).
9. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty476.
10. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
11. 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).
12. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
13. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
14. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**,

e1000529 (2009).

15. Ballardini, N. *et al.* IgE antibodies in relation to prevalence and multimorbidity of eczema, asthma, and rhinitis from birth to adolescence. *Allergy* **71**, 342–349 (2016).
16. Gref, A. *et al.* Genome-Wide Interaction Analysis of Air Pollution Exposure and Childhood Asthma with Functional Follow-up. *Am. J. Respir. Crit. Care Med.* **195**, 1373–1383 (2017).
17. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).
18. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
19. Roquer, J. *et al.* Acute stroke unit care and early neurological deterioration in ischemic stroke. *J. Neurol.* **255**, 1012–1017 (2008).
20. Adams, H. P. *et al.* Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* **24**, 35–41 (1993).
21. Soriano-Tárraga, C. *et al.* Epigenome-wide association study identifies TXNIP gene associated with type 2 diabetes mellitus and sustained hyperglycemia. *Hum. Mol. Genet.* **25**, 609–619 (2016).
22. Wright, J. *et al.* Cohort Profile: the Born in Bradford multi-ethnic family cohort study. *Int. J. Epidemiol.* **42**, 978–991 (2013).
23. Powell, J. E. *et al.* The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PLoS One* **7**, e35430 (2012).
24. McRae, A. F. *et al.* Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol.* **15**, R73 (2014).
25. Medland, S. E. *et al.* Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am. J. Hum. Genet.* **85**, 750–755 (2009).
26. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.*

48, 1284–1287 (2016).

27. Olsen, J. O. M. M. S. F., Olsen, J., Melbye, M. & Olsen, S. F. The Danish National Birth Cohort its background, structure and aim. *Scandinavian Journal of Public Health* vol. 29 300–307 (2001).
28. Paternoster, L. *et al.* Genome-wide population-based association study of extremely overweight young adults--the GOYA study. *PLoS One* **6**, e24303 (2011).
29. Poulton, R., Moffitt, T. E. & Silva, P. A. The Dunedin Multidisciplinary Health and Development Study: overview of the first 40 years, with an eye to the future. *Soc. Psychiatry Psychiatr. Epidemiol.* **50**, 679–693 (2015).
30. Bowtell, D. D. L. Rapid isolation of eukaryotic DNA. *Anal. Biochem.* **162**, 463–465 (1987).
31. Jeanpierre, M. A rapid method for the purification of DNA from blood. *Nucleic Acids Res.* **15**, 9611–9611 (1987).
32. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
33. Moffitt, T. E. & E-Risk Study Team. Teen-aged mothers in contemporary Britain. *J. Child Psychol. Psychiatry* **43**, 727–742 (2002).
34. Odgers, C. L., Caspi, A., Bates, C. J., Sampson, R. J. & Moffitt, T. E. Systematic social observation of children's neighborhoods using Google Street View: a reliable and cost-effective method. *J. Child Psychol. Psychiatry* **53**, 1009–1017 (2012).
35. Marzi, S. J. *et al.* Analysis of DNA Methylation in Young People: Limited Evidence for an Association Between Victimization Stress and Epigenetic Variation in Blood. *Am. J. Psychiatry* **175**, 517–529 (2018).
36. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).
37. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95–103 (1999).
38. The InterAct Consortium. Design and cohort description of the InterAct Project: an

examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. *Diabetologia* **54**, 2272–2282 (2011).

39. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
40. Bendinelli, B. *et al.* Fruit, vegetables, and olive oil and risk of coronary heart disease in Italian women: the EPICOR Study. *Am. J. Clin. Nutr.* **93**, 275–283 (2011).
41. Palli, D. *et al.* A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori* **89**, 586–593 (2003).
42. Guarrera, S. *et al.* Gene-specific DNA methylation profiles and LINE-1 hypomethylation are associated with myocardial infarction risk. *Clin. Epigenetics* **7**, 133 (2015).
43. Kaprio, J. The Finnish Twin Cohort Study: an update. *Twin Res. Hum. Genet.* **16**, 157–162 (2013).
44. Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31**, 1243–1264 (2016).
45. 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**, 37 (2015).
46. Strandberg, T. E., Järvenpää, A. L., Vanhanen, H. & McKeigue, P. M. Birth outcome in relation to licorice consumption during pregnancy. *Am. J. Epidemiol.* **153**, 1085–1088 (2001).
47. Price, E. M. & Robinson, W. P. Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Front. Genet.* **9**, (2018).
48. Guxens, M. *et al.* Cohort Profile: The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project. *Int. J. Epidemiol.* **41**, 930–940 (2011).
49. Arshad, S. H., Hasan Arshad, S., Karmaus, W., Zhang, H. & Holloway, J. W. Multigenerational cohorts in patients with asthma and allergy. *J. Allergy Clin. Immunol.* **139**, 415–421 (2017).

50. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16**, 1215 (1988).
51. Westendorp, R. G. J. *et al.* Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *J. Am. Geriatr. Soc.* **57**, 1634–1637 (2009).
52. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).
53. Deelen, J. *et al.* Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age. *Hum. Mol. Genet.* **23**, 4420–4432 (2014).
54. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1042r (2018).
55. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res.* **24**, 1725–1733 (2014).
56. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
57. Antoni, G. *et al.* Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Med. Genet.* **12**, 102 (2011).
58. Antoni, G. *et al.* A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. *J. Thromb. Haemost.* **8**, 2671–2679 (2010).
59. Oudot-Mellakh, T. *et al.* Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. *Br. J. Haematol.* **157**, 230–239 (2012).
60. Huang, J. *et al.* Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. *Blood* **120**, 4873–4881 (2012).
61. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).

62. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
63. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
64. Chen, Y.-A. *et al.* Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation. *American journal of human genetics* vol. 91 762–764 (2012).
65. Chen, Y.-A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
66. Germain, M. *et al.* Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS One* **7**, e38538 (2012).
67. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
68. Johnson, E. O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum. Genet.* **132**, 509–522 (2013).
69. Willemsen, G. *et al.* The Adult Netherlands Twin Register: twenty-five years of survey and biological data collection. *Twin Res. Hum. Genet.* **16**, 271–281 (2013).
70. Boomsma, D. I. *et al.* An Extended Twin-Pedigree Study of Neuroticism in the Netherlands Twin Register. *Behav. Genet.* **48**, 1–11 (2018).
71. Willemsen, G. *et al.* The Netherlands Twin Register biobank: a resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* **13**, 231–245 (2010).
72. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115 (2016).
73. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur. J.*

Hum. Genet. **22**, 221–227 (2014).

74. Roshyara, N. R., Kirsten, H., Horn, K., Ahnert, P. & Scholz, M. Impact of pre-imputation SNP-filtering on genotype imputation results. *BMC Genet.* **15**, 88 (2014).
75. Wijga, A. H. *et al.* Cohort profile: the prevention and incidence of asthma and mite allergy (PIAMA) birth cohort. *Int. J. Epidemiol.* **43**, 527–535 (2014).
76. Hofmann-Apitius, M., Alarcón-Riquelme, M. E., Chamberlain, C. & McHale, D. Towards the taxonomy of human disease. *Nat. Rev. Drug Discov.* **14**, 75–76 (2015).
77. Teruel, M., Chamberlain, C. & Alarcón-Riquelme, M. E. Omics studies: their use in diagnosis and reclassification of SLE and other systemic autoimmune diseases. *Rheumatology* **56**, i78–i87 (2017).
78. Girchenko, P. *et al.* Prediction and Prevention of Preeclampsia and Intrauterine Growth Restriction (PREDO) study. *Int. J. Epidemiol.* dyw154 (2016).
79. - Project MinE Consortium *et al.* Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. (2017) doi:10.1101/152553.
80. Newnham, J. P., Evans, S. F., Michael, C. A., Stanley, F. J. & Landau, L. I. Effects of frequent ultrasound during pregnancy: a randomised controlled trial. *Lancet* **342**, 887–891 (1993).
81. Fortin, J.-P., Fertig, E. & Hansen, K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Res.* **3**, 175 (2014).
82. van Iterson, M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics* **30**, 3435–3437 (2014).
83. Assenov, Y. *et al.* Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
84. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
85. Ikram, M. A. *et al.* The Rotterdam Study: 2018 update on objectives, design and main results. *Eur. J. Epidemiol.* **32**, 807–850 (2017).

86. Tillin, T., Forouhi, N. G., McKeigue, P. M. & Chaturvedi, N. Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. *Int. J. Epidemiol.* **41**, 33–42 (2010).
87. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
88. Datta, S. R. *et al.* A threonine to isoleucine missense mutation in the pericentriolar material 1 gene is strongly associated with schizophrenia. *Mol. Psychiatry* **15**, 615–628 (2010).
89. Hannon, E. *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* **17**, 176 (2016).
90. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
91. Pausova, Z. *et al.* Genes, maternal smoking, and the offspring brain and body during adolescence: design of the Saguenay Youth Study. *Hum. Brain Mapp.* **28**, 502–518 (2007).
92. Pausova, Z. *et al.* Cohort Profile: The Saguenay Youth Study (SYS). *Int. J. Epidemiol.* **46**, e19 (2017).
93. Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res. Hum. Genet.* **16**, 144–149 (2013).
94. Kurushima, Y. *et al.* Epigenetic findings in periodontitis in UK twins: a cross-sectional study. *Clin. Epigenetics* **11**, 27 (2019).
95. Beaumont, M. *et al.* Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biol.* **17**, 189 (2016).
96. UK Data Service › DOI. http://doi.org/10.5255/UKDA-SN-6614-12.
97. Gorrie-Stone, T. J. *et al.* Bigmelon: tools for analysing large DNA methylation datasets. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty713.

98. Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Sci. Rep.* **7**, 11008 (2017).
99. 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 (2012).
100. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
101. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med. Genet.* **7**, 74 (2006).
102. Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* vol. 9 (2017).
103. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).
104. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).
105. Reynolds, L. M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nature Communications* vol. 5 (2014).
106. Slieker, R. C. *et al.* Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics Chromatin* **6**, 26 (2013).
107. Spiers, H. *et al.* Methylomic trajectories across human fetal brain development. *Genome Research* vol. 25 338–352 (2015).
108. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nature Neuroscience* vol. 19 48–54 (2016).
109. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
110. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158

(2011).

111. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

112. Moayyeri, A., Hammond, C. J., Valdes, A. M. & Spector, T. D. Cohort Profile: TwinsUK and healthy ageing twin study. *Int. J. Epidemiol.* **42**, 76–85 (2013).

113. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).