(Epi) genetics and twins

van Dongen, J.

2015

Appendix 5

Supplement to Chapter 6

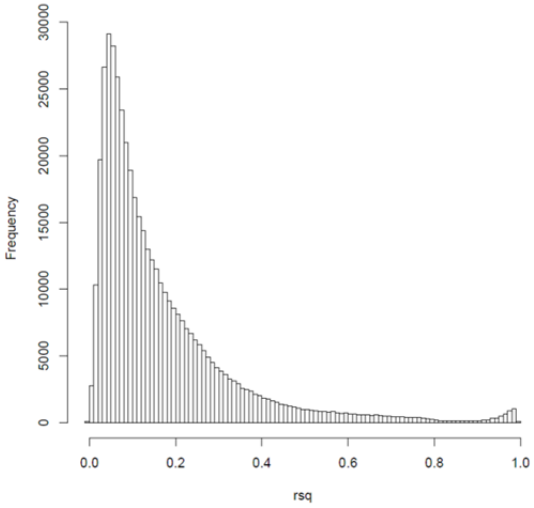
based on: Jenny van Dongen [*], Bastiaan T. Heijmans[*], Michel G. Nivard[*], Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V. Dolan, Erik A. Ehli, Gareth Davies, BIOS Consortium[v], H. Eka Suchiman, Rick Jansen, Joyce B. van Meurs, P. Eline Slagboom, Dorret I. Boomsma. The heritability of DNA methylation in peripheral blood: influences of common SNPs and variability of genetic and environmental variance with age and sex. (*manuscript in preparation*)

Supplementary Table 1: Twin correlations, heritability and longitudinal correlation, stratified by the amount of variation in DNA methylation between individuals.

		***r* MZ twins**			***r* DZ twins**			**Classical Twin Heritability**			**Longitudinal Correlation**		
SD of the β-value	N CpGs	mean	SD	median	mean	SD	median	mean	SD	median	mean	SD	median
>= 0 (All)	411169	0.20	0.21	0.13	0.09	0.11	0.06	0.22	0.27	0.16	0.21	0.30	0.16
>= 0-0.01	49599	0.04	0.05	0.04	0.02	0.05	0.02	0.05	0.13	0.04	0.01	0.19	0.01
>= 0.01-0.02	171461	0.09	0.09	0.08	0.04	0.07	0.04	0.10	0.17	0.09	0.07	0.20	0.07
>= 0.02-0.03	85003	0.23	0.15	0.23	0.11	0.09	0.10	0.25	0.22	0.24	0.24	0.26	0.24
>= 0.03-0.04	49918	0.35	0.19	0.36	0.15	0.11	0.15	0.38	0.25	0.38	0.39	0.28	0.43
>= 0.04-0.05	27110	0.44	0.20	0.44	0.19	0.12	0.19	0.49	0.28	0.52	0.51	0.29	0.57
>= 0.05	28078	0.57	0.25	0.60	0.25	0.13	0.25	0.64	0.33	0.73	0.64	0.31	0.77

SD= Standard deviation. β-value= Methylation beta-value, which represents the proportion of DNA Methylation. The first row summarizes the results for all analyzed CpGs and all other rows summarize the results for CpGs grouped based on the standard deviation of methylation level across all subjects.

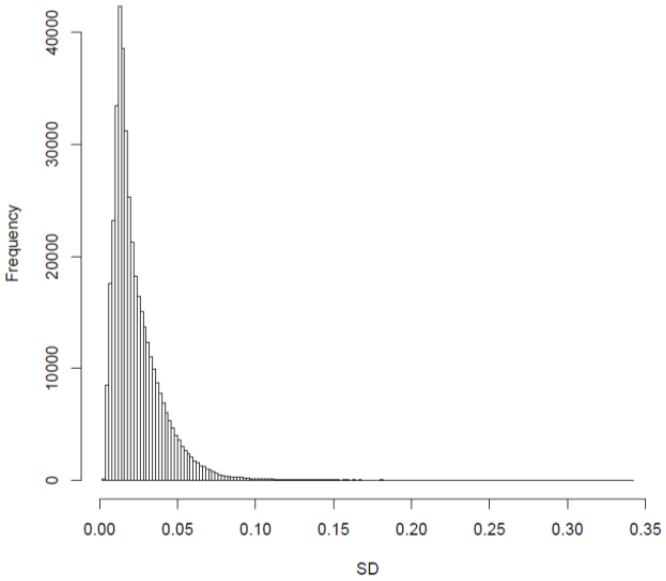
Figure S1: Histogram of the variance of DNA methylation level explained by covariates at individual genome-wide CpGs.



[v] The Biobank-based Integrative Omics Study (BIOS) Consortium

Rsq=Adjusted r-squared from a linear regression model with DNA methylation level at one CpG site as outcome and the following predictors: sex, age, array row, 96-wells plate (dummy coded), white blood cell percentages (neutrophils, monocytes and eosinophils; assessed at sample collection), and the first ten PCs derived from the genotype data.

Figure S2: Histogram of the standard deviation of the methylation β-value for genome-wide CpGs.



SD=Standard Deviation of DNA methylation level (β-value).

Figure S3: Histogram of the difference in heritability between females and males (h^2 in females minus h^2 in males) for 2654 CpGs with significant interaction between sex and genetic variance or between sex and unique environmental variance.

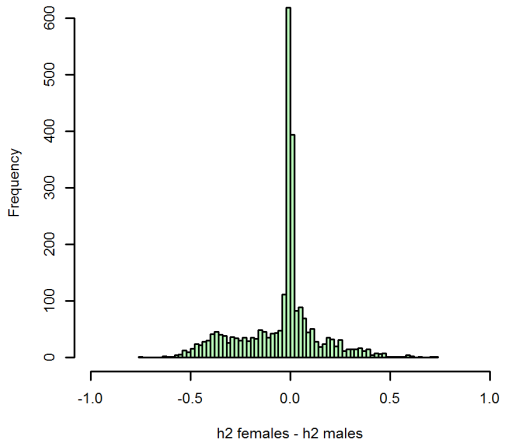
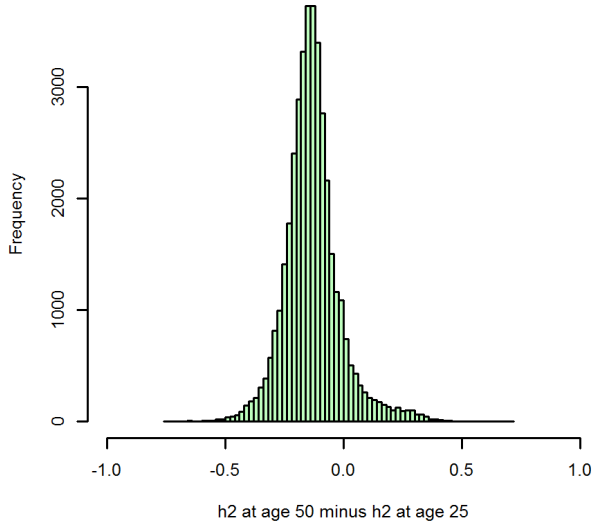


Figure S4: Histogram of the difference in (IBD-based) heritability between age 50 and age 25 (h^2 at age 50 minus h^2 at age 25) for 39194 CpGs with significant interaction between age and genetic variance or between age and unique environmental variance.



Supplementary Methods

Genome-wide SNP data

Three distinct genotype datasets were available. The first consisted of previously collected genome-wide SNP data that were only used as part of the Quality Control (QC) procedure of the DNA methylation data. The second previously collected genome-wide SNP data were used only as part of the statistical analyses of the DNA methylation data. The third SNP dataset consisted of 65 common SNPs targeted by the Illumina 450k array that were only used as part of QC procedure of the DNA methylation data.

Genotype data used during QC of the DNA methylation data

Of the 3221 subjects for whom peripheral blood methylation samples were assessed with the illumina 450k array, 2665 subjects had been previously genotyped or had a MZ co-twin who had been genotyped one or multiple times on any of the following genotype arrays: Affymetrix6, Affymetrix5, and Illumina660. One set of genotypes was selected (the one with the best quality) for MZ twins if both twins were genotyped and for individuals who had been genotyped on multiple platforms. In total, 1870 genome-wide SNP data sets were available, which were informative 2665 individuals (including 795 MZ co-twins). For the DNA methylation data QC, the overlapping SNPs from the Affymetrix6, Affymetrix5, and Illumina660 arrays were selected. Because of the small overlap of SNPs on these three arrays, this dataset was not used for the heritability analyses of DNA methylation

Genotype data used in the heritability analyses of DNA methylation

The analyses of DNA methylation heritability were performed using genome-wide SNP data collected with the Affymetrix6 array and SNP data that were extracted from whole genome sequence data that were available for a small subset of subjects (described previously)[1]. Of the 2975 subjects with good quality DNA methylation data and data on white blood cell counts, Affy6 genotype data were available for 2289 subjects and sequence data for 341 individuals (numbers include both MZ twins). Only SNPs present on the Affy6 platform were extracted from the sequence data. For a subset of 84 subjects for whom sequence data and Affy6 data were available, the sequence data were selected. SNPs with an allele frequency difference between individuals genotyped on Affy6 and individuals who were sequenced were removed (based on a p-value < $1*10^{-5}$ in a case-control genome-wide association analysis, where case-control status reflected whether a person was genotyped on Affy6 or whole-genome sequenced). The genome-wide SNP data were used to construct a genetic relatedness matrix (GRM), which summarizes overall genetic relatedness between all subjects (N=2603) based on all genotyped autosomal SNPs (MAF > 0.01) with Genome-wide Complex Trait Analysis (GCTA)[2].

DNA methylation Quality Control and filtering of methylation probes

Quality control and processing of the DNA methylation data from buccal samples has been previously described[3]. The following text describes the quality control and processing of the DNA methylation data from blood samples. The raw intensity files (idat) were imported into the R environment[4], where further processing, quality control and normalization took place using a protocol developed by the LUMC Molecular Epidemiology department.

First, the methylation data were examined with the R package MethylAid[5], which marks outlier samples for a number of quality metrics that are computed based on sample dependent and sample independent quality metrics. The performance of the 3264 samples is plotted for each of five quality metric in Supplementary Figure 5-9. Only samples that passed all five quality criteria (using the default MethylAid thresholds) were kept for further analyses. In total, 70 low-performing samples were excluded (2.1%), the majority of which failed based on multiple criteria (Supplemental Table 2). Only the 3194 samples showing good overall quality were taken on to further processing steps.

Several probe-level QC steps were performed to filter out probes with low performance. For all samples, ambiguously-mapped probes were excluded, based on the definition of an overlap of at least 47 bases per probe from Chen et al[6], and all probes containing a SNP, identified in the Dutch population[1], within the CpG site (at the C or G position) were excluded, irrespective of minor allele frequency. For each sample individually, probes with an intensity value of zero (not present on the array of a particular sample), probes with a detection P value > 0.01 (calculated using the function detectionP from the minfi package[7]), and probes with a bead count < 3 were excluded. After these steps, probes with a success rate < 0.95 across samples were removed from all samples and the success rate across probes for each sample was computed (Mean per sample success rate=99.89%, range=97.86%-99.96%). The total number of CpGs after these filtering steps was 421119. Only autosomal sites were kept in the current analyses (N=411169).

We performed several checks to confirm sample identity, by making use of previously collected genotype data, 65 SNP (control) probes targeted by the 450k array,

and differential methylation patterns in males versus females. Previously collected raw genotype data was used as input for the program MixupMapper, which computes the probability that a DNA methylation sample matches supplied genotype information based on mQTLs estimated from the dataset[8]. To confirm sex, we clustered samples based on their methylation data, by calculating the Euclidean distance from the pair-wise correlations between samples followed by hierarchical clustering (cluster method=complete linkage). Clustering based on all probes and based on probes on the sex chromosomes only yielded similar results. We computed the correlation between samples for 65 SNP (control) probes targeted by the 450k array to confirm zygosity of twins, and to confirm that longitudinal samples indeed belonged to the same person. Finally, we used the 65 SNP probes to examine potential contamination of samples with foreign DNA, by computing the number of SNPs per sample with an unclear genotype (which we defined as SNPs where the proportion of signal from each allele lay between 0.2 and 0.4 or between 0.6 and 0.8, on a scale from o to 1, i.e. a pattern not clearly supporting membership to any of the three genotype classes. The number of 'unclear genotypes' showed a mean of 3.3 across all samples (median=2, SD=3.5, Supplemental Figure 10). We excluded samples with >= 15 unclear genotypes (99th percentile). The genome-wide methylation distribution of these excluded samples showed relatively more intermediate methylation levels (Supplemental Figure 11). An example scatterplot of the 65 SNP probes in MZ twin samples illustrating DNA contamination of the sample of one of the twins, as detected by this method, is given in Supplementary Figure 12.

In total, 132 samples were involved in at least one of the following issues: genotype mismatch, sex-mismatch, DNA contamination, inconsistent SNP probe correlation (either between twins or between longitudinal samples from the same person). After solving a swap between 2 methylation samples identified by MixupMapper (and confirmed by the other checks) by re-swapping methylation data IDs (leaving 128 samples with issues), 67 samples were excluded based on the following grounds: only sex mismatch (22 samples), only genotype mismatch (10 samples), only DNA contamination (27 samples), genotype + sex mismatch (6 samples), DNA contamination + sex mismatch (2 samples). After removal of these samples, there were still 38 samples with an inconsistent SNP probe correlation (involving i.e. a zygosity mismatch or mismatch between longitudinal samples), which were all excluded, giving a total of 105 samples (3.3%) excluded based on failed identity or contamination, on top of the 70 samples excluded based on bad quality of the methylation data.

Finally, for 22 persons with 450k methylation data available from blood and buccal samples, the 65 SNP probes confirmed that blood and buccal samples indeed belonged to the same individual.

Exploration of technical and biological confounding

To get an impression of the impact of technical and biological effects on overall variation in methylation, Principle Component Analysis (PCA) was performed on the raw genome-wide methylation data (Supplemental Table 3, Supplemental Figure 13), and the correlation between PC scores and several known technical batches and biological outcomes were computed. PC1 related to sex (r=0.92), PC2 was strongly correlated with position on the array (in particular, array row, r=0.50), PC3 with several white blood cell counts (e.g. lymphocytes: r=0.45), and PC4 with age (r=-0.59). Other batch variables (e.g. 96-wells plate, array, scanner) correlated to a smaller degree with multiple components.

Normalization of the methylation data and correction for covariates

To reduce technical variability between samples while retaining as much biological variation in DNA methylation as possible, the data were normalized with Functional Normalization, a between-sample normalization method that normalizes the data using PCs (the number of which is user-specified) estimated from control probes that are specifically designed not to measure biological variation in samples[9]. We chose to perform Functional Normalization with the first 4 PCs, because PCA based on the data from control probes showed that in our data only the first four PCs had an eigen value > 1 (Supplemental Table 4).

Normalized intensity values were converted into beta-values (β) and M-values[10]; β -values were used for descriptive purposes only because of their biological interpretability, while M-values were used as input for all analyses. The β -value, which represents the methylation level at a CpG for an individual and ranges from 0 to 1, is calculated as:

$\beta = M/(M+U+\alpha)$

where M=Methylated signal, U=Unmethylated signal, and α represents a correction term (100 by default) to control the β –value of probes with very low overall signal intensity (i.e. probes for which M+U~0 after normalization).

The M-value is equivalent to a log2 logistic transformation of β:

$M=\log2(M+\alpha/U+\alpha) = \log2(\beta/1- \beta)$.

Supplemental Table 2: Number of bad quality DNA methylation samples that failed sample quality checks.

Quality Metric	**N outliers[A]**
MU	47
BS	52
NP	44
HC	22
DP	48
Combinations of failure	**N outliers[B]**
BS	1
HC	14
HC + BS	5
HC + BS + MU	2
DP	3
DP + DS + MU	1
DP + NP + MU	1
DP + NP + MU + BS + MU	42
DP + NP + MU + BS + MU + HC	1
Total N bad quality samples[C]	70

The following five quality metrics were computed with the R package MethylAid: **MU=** median Methylated versus Unmethylated signal intensity, **BS=**Efficiency of bisulphite conversion, **NP=**overall quality based on sample-dependent control probes (non-polymorphic quality control probes), **HC=** overall quality based on sample-independent hybridization control probes. **DP=** Fraction of probes per sample where the signal exceeds the background signal, as assessed with the detection p-value, which uses the negative control probes to assess background signal.

[A]N outliers= Number of samples that failed based on each quality metric.

[B]N outliers= Number of samples that failed based on a particular combination of multiple quality metrics.

[C]All samples that failed based on one or more of the five quality metrics were discarded (70 samples).

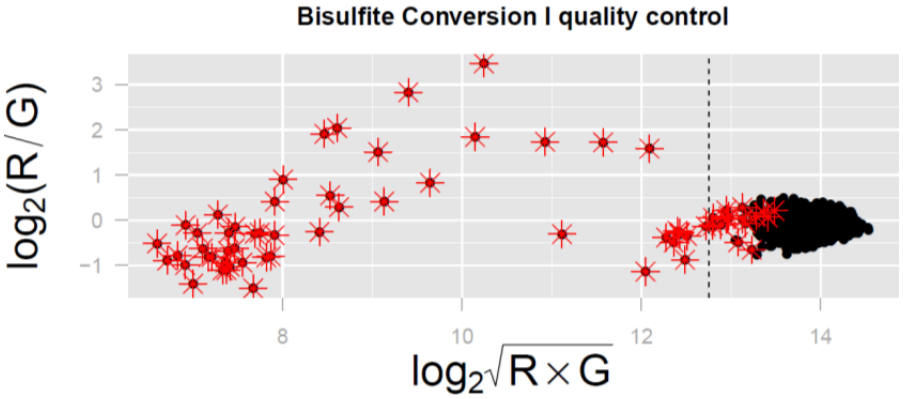
Supplemental Table 3: Eigen values and proportion of variance explained by Principle Components 1 to 15, calculated based on the raw genome-wide methylation data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Eigen value	54.71	37.13	14.09	5.65	3.95	2.71	1.51	1.23	0.98	0.95	0.68	0.65	0.55	0.49	0.46
Proportion of Variance	0.2211	0.1501	0.0569	0.0228	0.0160	0.0110	0.0061	0.0050	0.0040	0.0038	0.0028	0.0026	0.0022	0.0020	0.0019
Cumulative Proportion	0.2211	0.3712	0.4281	0.4510	0.4669	0.4779	0.4840	0.4890	0.4929	0.4968	0.4995	0.5021	0.5044	0.5063	0.5082

Supplemental Table 4: Eigen values and proportion of variance explained by Principle Components 1 to 10, calculated based on control probes from the methylation array.

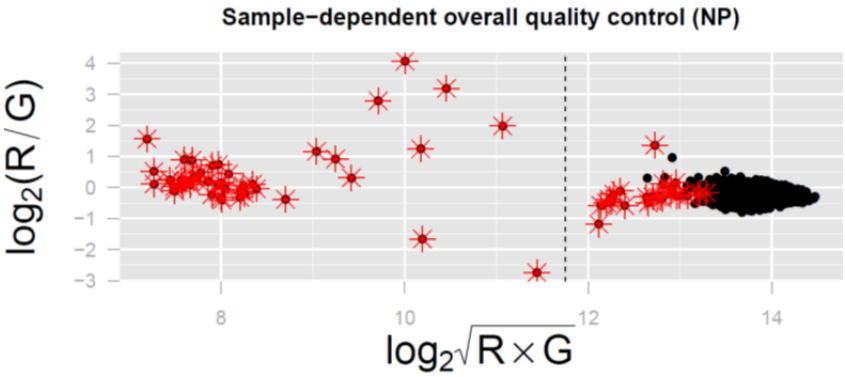
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Eigen value	25.37	7.09	3.61	1.36	0.73	0.48	0.45	0.36	0.30	0.25
Proportion of Variance	0.6039	0.1687	0.0860	0.0323	0.0175	0.0114	0.0108	0.0087	0.0071	0.0060
Cumulative Proportion	0.6039	0.7727	0.8587	0.8910	0.9085	0.9199	0.9307	0.9393	0.9465	0.9524

Figure S5: Quality control plot of bisulfite conversion.



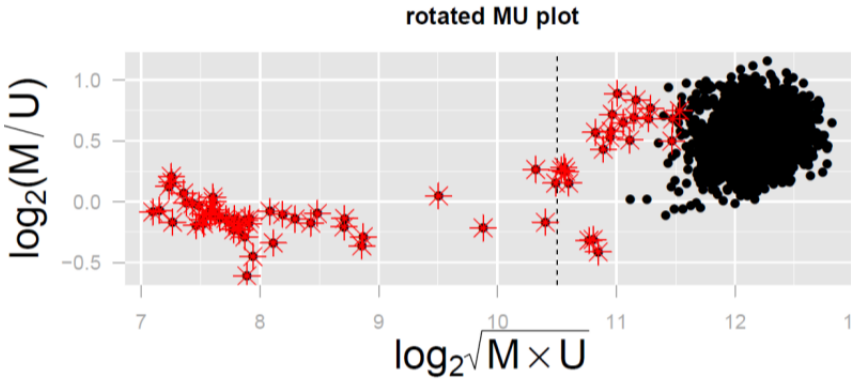
The performance of bisulfite conversion quality control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics. R=Red Channel. G=Green Channel.

Figure S6: Quality control plot of overall sample quality based on sample-dependent control probes (Non-Polymorphic quality control probes)



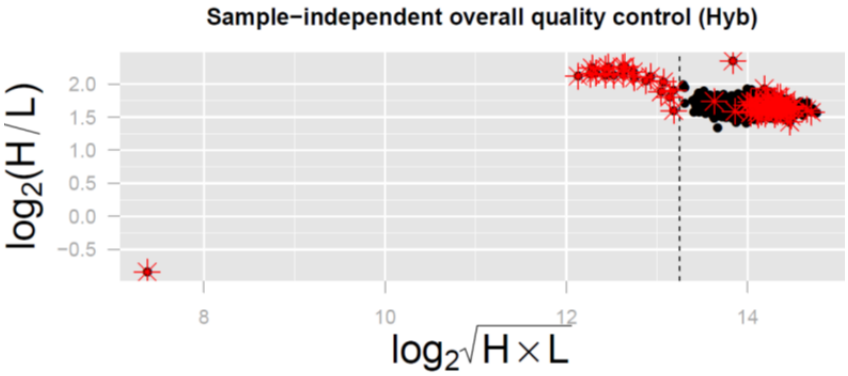
The performance of Non-Polymorphic quality control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics. R=Red Channel. G=Green Channel.

Figure S7: Quality control plot of the median Methylated versus Unmethylated signal intensity.



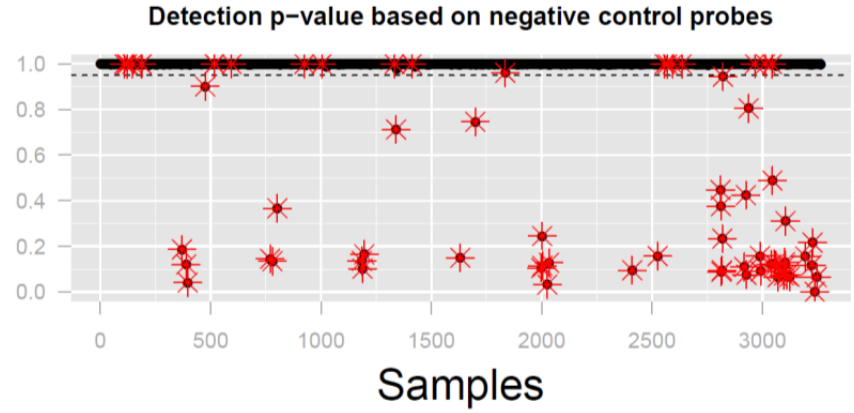
The relationship between the Median Methylated (M) and Unmethylated (U) signal intensity is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S8: Quality control plot based on sample-independent hybridization control probes.



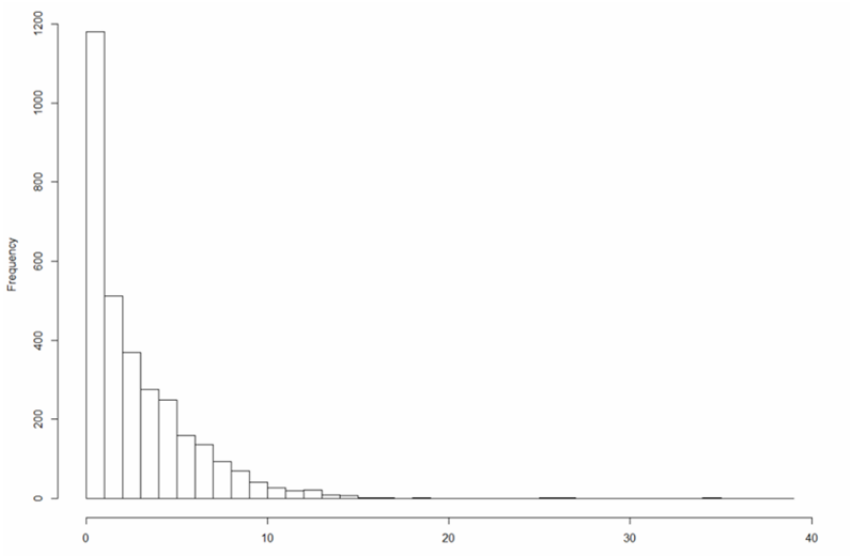
The performance of sample-independent hybridization control probes is plotted for all DNA methylation samples. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S9: Quality control plot showing the proportion of probes with a detection p-value < 0.01 within samples.



For all methylation samples, the proportion of probes per sample with a detection p-value < 0.01 is plotted (y-axis). The detection p-value indicates whether the probe signal exceeds the background signal, where the background signal is calculated using the negative control probes. Red stars denote samples that failed on the basis of any of the five quality metrics.

Figure S10: Histogram of the number of Illumina 450k SNP probes per sample displaying an unclear genotype.



X-axis= the number of unclear genotype per sample: SNPs where the proportion of signal from each allele lay between 0.2 and 0.4 or between 0.6 and 0.8, on a scale from o to 1, i.e. a pattern not clearly supporting membership to any of the three genotype classes. In total 65 common SNPs from the Illumina 450k array were assessed. Y-axis=Number of methylation samples. Methylation samples with >= 15 unclear genotypes (99th percentile) were excluded from analyses.

Figure S11: DNA methylation density plot showing samples excluded based on suspected DNA contamination (1/orange) and all other samples (0/green).

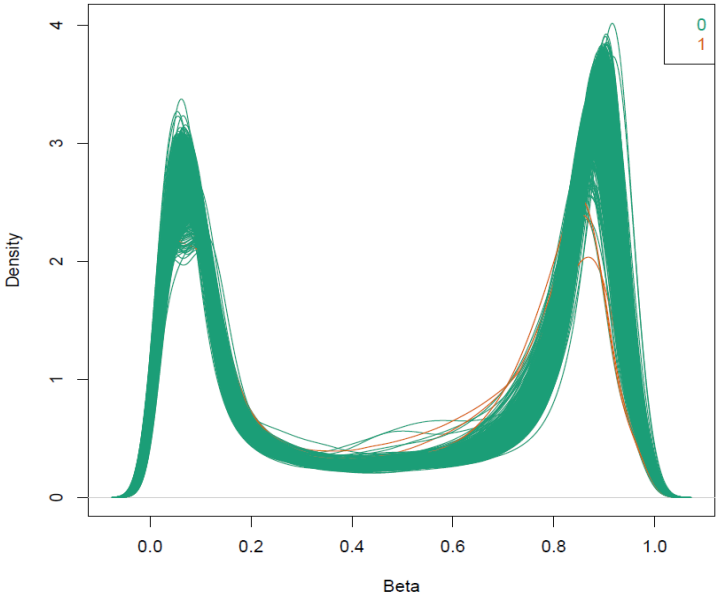


Figure S12: Example scatterplot of the 65 Illumina 450k SNP probes in one pair of MZ twins, of which one twin (on the y-axis) was excluded based on suspected DNA contamination.

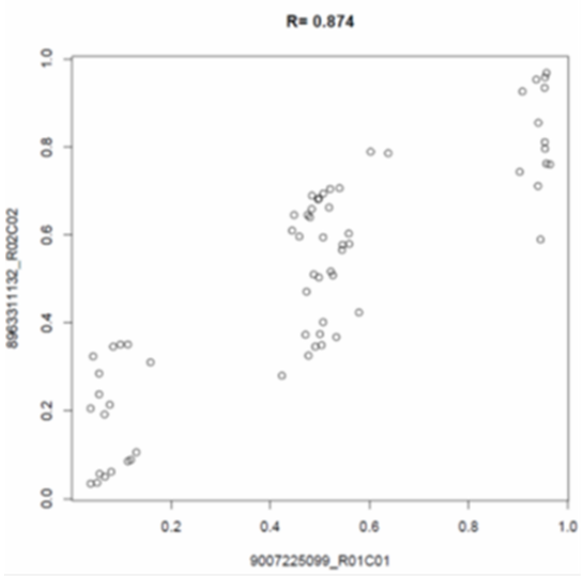
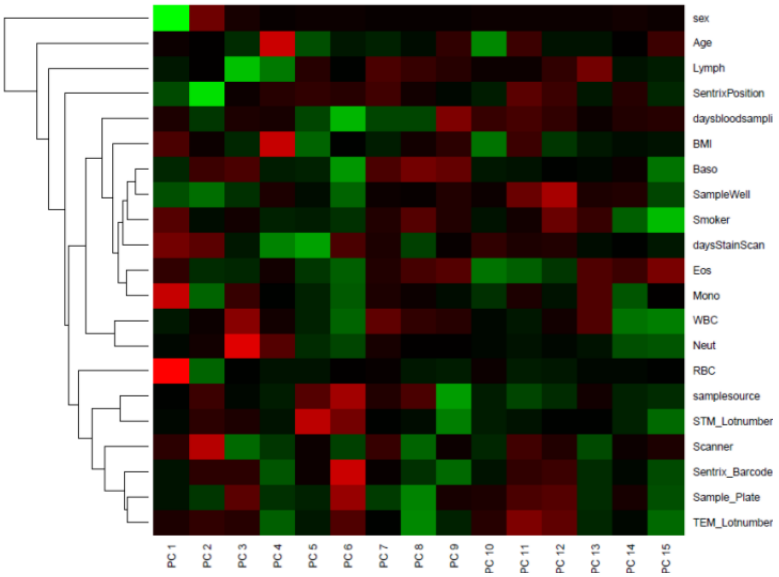


Figure S13: Heatmap depicting the correlation between the first 15 Principal Components from the raw genome-wide methylation data (x-axis) and technical batches and biological effects (y-axis).



Stronger green=larger positive correlation. Stronger red= larger negative correlation. Lymph=lymphocyte counts. Daysbloodsmpli=Days between blood sampling and hybridization. BMI=Body Mass Index. Baso=Basophil count. daysStainScan=Days

between staining and scanning. Eos=Eosinophil count. Mono=Monocyte count. WBC=Total white blood cell count. Neut=Neutrophil count. RBC=Red Blood cell count. Samplesource= 1 or 2, for individuals with 2 longitudinal samples (both from blood).

Reference List

1. Genome of the Netherlands Consortium Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics* **46**, 818-825 (2014).
2. Yang,J., Lee,S.H., Goddard,M.E., & Visscher,P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76-82 (2011).
3. van Dongen,J. *et al.* Epigenetic variation in monozygotic twins: a genome-wide analysis of DNA methylation in buccal cells. *Genes (Basel)* **5**, 347-365 (2014).
4. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. 2013.
5. van Iterson M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics.*(2014).
6. Chen,Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* **8**, 203-209 (2013).
7. 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).
8. Westra,H.J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics.* **27**, 2104-2111 (2011).
9. Fortin,J.P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *bioRxiv*(2014).
10. Du,P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC. Bioinformatics.* **11**, 587 (2010).