Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution

Maarten van Iterson[1], Erik van Zwet[2], the BIOS Consortium*, P. Eline Slagboom[1] and Bastiaan T. Heijmans[1]

[1]Department of Molecular Epidemiology, [2]Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands *the BIOS Consortium (**Supplemental Text**).

Correspondence to:
M. van Iterson
Leiden University Medical Center, Department of Molecular Epidemiology, Leiden, The Netherlands
m.van_iterson@lumc.nl
T: +31 (0) 71 526 9730

Supplementary Methods

Data sets DNA methylation data and RNA-seq data were generated within the Biobank-based Integrative Omics Studies Consortium. The data comprises four biobanks: Cohort on Diabetes and Atherosclerosis Maastricht (CODAM)[40], LifeLines (LL)[41], Leiden Longevity Study (LLS)[42], the Rotterdam Study (RS)[43]. Sample identity of DNA methylation and expression data was confirmed using genotype data. Both RNA-seq fastq-files and DNA methylation idat-files are available from EGA (EGAC00001000277). Data was generated by the Human Genotyping facility (HugeF) of ErasmusMC, the Netherlands (www.glimDNA.org).

RNAseq data preprocessing Detailed description of the RNA-seq data processing can be found in Zhernakova *et al.*[20] Subsequently, RNA-seq counts were normalized using TMM[44] and transformed to log2 counts per million. Genes that yielded zero counts for all samples across cohorts were removed which resulted in 45867 genes (ENSEMBLv73). For all analyses genes with the lowest overall variance were excluded (5% lowest).

450k DNA methylation data preprocessing Sample quality control was performed using MethylAid[45]. Ambiguously mapped probes[46], probes with a high detection P-value (> 0.01), probes with a low bead count (< 3 beads) and probes with a low success rate (missing in >95% of the samples) were set to missing. Samples containing an excess of missing probes (> 5%) were excluded from the analysis. Subsequently, per cohort, imputation[47] was performed to impute the missing values. Functional normalization[48], as implemented in the minfi-package[49], was used per cohort. All analyses where performed on M-values (detailed description of the 450K DNA methylation preprocessing steps are available at: https://git.lumc.nl/molepi/Leiden450K).

White blood cell count prediction White blood cell counts (WBC), i.e., neutrophils, lymphocytes, monocytes, eosinophils and basophils, were measures by the standard WBC differential as part of the CBC (Complete Blood Count). However, a minority of samples were lacking CBC measurements. Since DNA methylation levels are informative of the white blood cell composition[50] we build a linear predictor to infer the white blood cell composition of those samples lacking WBC measurements (**Supplemental Text**). In the meta-analyses all samples where used, also those lacking cell counts, only for this part of the manuscript predicted/imputed cell counts were used.

Association Analyses All association analyses where performed using limma's lmFit-function[51]. Since, the sample sizes of our data were all above >100 we bypassed the empirical Bayes step. Furthermore, test-statistics where transformed to *P* values using a standard normal distribution. For the association analyses on RNA-seq data we first applied a voom-transformation[35] on the TMM-normalized counts while controlling for known covariates including age, gender, smoking status measured cell counts, flow-cell.
For the association analyses on DNA methylation data the functional normalized beta-values were

transformed to M-values and again lmFit from limma was used to obtain test-statistics for the covariate of interest. Here we included, age, gender, smoking status, measured cell counts and array position as known covariates.

Estimation of the unobserved covariates To investigate if adding estimated unobserved covariates reduces bias and inflation we performed association analyses with only the covariate of interest, known covariates (e.g. white blood cell counts) and either one or three principal components, estimated from the data, and CATE[24] for both the EWAS and TWAS and specifically for the TWAS we used additionally RUV[23,29] and SVA[22] and for the EWAS ISVA[30] and RUVm[52].

Simulation studies
The impact of true association on the genomic inflation factor Hundred sets of 2000 test-statistics were generate from a normal mixture distribution with different mixture coefficients (0.8, 0.90 and 0.95). The majority of the null test-statistics were drawn from a standard normal, N(0, 1), while the alternative test-statistics were drawn from a normal distribution, $\mathcal{N}(\mu, 1)$, with $\mu \sim \mathcal{N}(0, 3)$. A scenario was used with an equally number of positive and negative associations. For each set of test-statistics inflation factors were calculated to investigate the impact of the amount of true association.
Comparing different methods that estimate the empirical null distribution A few methods have been proposed to estimate an empirical null distribution for a set of test-statistics. In order to compare there performance sets of test-statistics were generated under different scenario's; scenario equal containing equal number of positive and negative associations (0.05, 0.05), scenario skewed containing only positive associations (0.1), scenario small similar to scnerio equal with only 0.01, scenario close where the distribution for the means had expected value of 1 (in stead of 3). For each scenario 2000 test-statistics where generated 100 times. To estimate an empirical null distribution we used the locfdr[16] R-package both the maximum-likelihood and moment-matching method with default parameters and our novel Bayesian approach.
Simulation with unobserved covariates We used the same simulation setup to generated unobserved covariates as implemented in the R package cate using the gen.sim.data-function.

The Gibbs Sampler A Gibbs sampling algorithm[26,53,54] is used to obtain samples from the joint distribution of the three component normal mixture with 9-1 parameters (minus one, since the mixture proportions are constraint to sum to one). Standard conjugate priors are used for the means, μ_j, $j = 1, 2, 3$, variances, σ_j^2, and mixture proportions, ϵ_j. Hence, we assume for $\mu_j|\sigma_j^2 \sim \mathcal{N}(\lambda_j, \sigma_j^2/\tau_j)$, $\sigma_j^2 \sim \mathrm{IG}(\alpha_j, \beta_j)$, and the mixture proportion prior distribution is a Dirichlet distribution $\epsilon_j \sim \mathcal{D}(\gamma_j)$. Well chosen hyperpriors ensure that labeling switching, i.e., during sampling from the posterior, the null component is switched with one of the alternative components, does not occur easily. That is, we take hyperpriors for μ_j for the null component $\lambda_0 = 0$ and $\tau_0 = 1000$ and $\lambda_\pm = \pm 3$ and $\tau_\pm = 100$, for the alternative components. The hyperpriors for the variance parameters are equal for all components $\alpha = 1.28$ and $\beta = 0.36\mathrm{mad}(t_1, \cdots, t_p)$. Since, we known in advance that the majority (>80%) of the test-statistics will follow the null distribution “informative” priors for the mixture proportions are used as well, namely $\gamma_0 = 90$, $\gamma_\pm = 5$. Such that the prior Dirichlet distribution has expected values 0.9 and 0.05 with variances 0.03 and 0.02 for the null and alternative components. Furthermore, data-dependent starting values are used to start the algorithm at good initial point (actually, these are the mad and median estimates of bias and inflation). We use a burnin-period of 3000 iterations and use 2000 subsequent samples to estimate the parameters of the mixture distribution using the mean. The Gibbs Sampler algorithm is implemented in C and and uses a fast sampling approach for generating samples from the multinomial distribution[55].
Optionally, test-statistics other then normal can be use, e.g., chi-square statistics, by first applying Efron's z-transformation, e.g., qnorm(pchisq(t,df))and use the corresponding z-scores as input.
The algorithm is as follows: Given test-statistics (z-scores or transformed to z-scores) y_i for $i = 1, \cdots, p$, prior distributions with hyper-parameters,

$$\epsilon_j \sim \mathcal{D}(\gamma_j), \quad \mu_j|\sigma_j^2 \sim N(\lambda, \tau), \quad \sigma_j^2 \sim IG(\alpha, \beta)$$

and starting values for the posterior distributions.

Iterate for $t = 1, \cdots, 5000$,
1) generate the missing (unobserved) data: $z_{ij} \sim \mathcal{M}(\tilde{p}_{ij})$ from a multinomial distribution, with parameter $p_{ij} = \epsilon_j \phi(y_i; \mu_j, \sigma_j)$, $\tilde{p}_{ij}$ represents the normalized proportion $\sum_j \tilde{p}_{ij} = 1$.

2) generate samples from the posteriors:

$$\epsilon_j \sim \mathcal{D}(\gamma_j + n_j),$$

$$\mu_j|\sigma_j^2 \sim N\left(\frac{\lambda_j \tau_j + s_j}{n_j + \tau_j}, \frac{\sigma_j^2 + s_j}{n_j + \tau_j}\right),$$

$$\sigma_j^{-2} \sim \Gamma\left(\alpha + \frac{1}{2}(n_j + 1), \left(\beta + \frac{1}{2}\tau_j(\mu_j - \lambda_j)^2 + \frac{1}{2}s_j^2\right)^{-1}\right),$$

the latter mimics sampling from an inverse gamma distribution. For clarity, an iteration superscript is omitted. We assume 3000 iterations (burn-in period) is long enough for the Markov properties to hold such that the samples from the conditional distributions can be assumed to be samples from the joint parameter distribution.

REFERENCES Supplementary Methods

1. van Greevenbroek, M.M. *et al*. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.*, **41**, 4 (2011).
2. Westendorp, R.G. *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**, 9 (2009).
3. Tigchelaar, E.F. *et al*. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*, **5**, 8 (2015).
4. Hofman, A. *et al*. The Rotterdam Study: 2012 objectives and design update. *Eur. J. Epidemiol.*, **26**, 8 (2011).
5. Robinson, M.D. and Oshlack., A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology,* **11**, 3 (2010).
6. van Iterson, M. *et al*. MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics,* **30**, 23 (2014).
7. Chen, Y.-a. *et al*. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics,* ***8***, (2013).
8. Troyanskaya O, *et al*. Missing value estimation methods for DNA microarrays. *Bioinformatics,* **17**, 6 (2001).
9. Fortin, J.P. *et al*. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology,* **15**, 12 (2014).
10. Aryee, M.J. *et al*. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics,* **15**, 10 (2014).
11. Houseman, E.A. *et al*. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics,* **8**, 13 (2012).
12. Ritchie, M.E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.,* **20**, 43 (2015).
13. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 32, **2** (2016).
14. Geman, S. and Geman, D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian

Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 6 (1984).
15. Casella, G. and George, E.I. Explaining the Gibbs Sampler. *The American Statistician*, **46**, 3 (1992).
16. Efraimidisa, P.S. Spirakisb, P.G. Weighted random sampling with a reservoir. *Information Processing Letters*, 97, **6** (2006).

Supplementary Figures

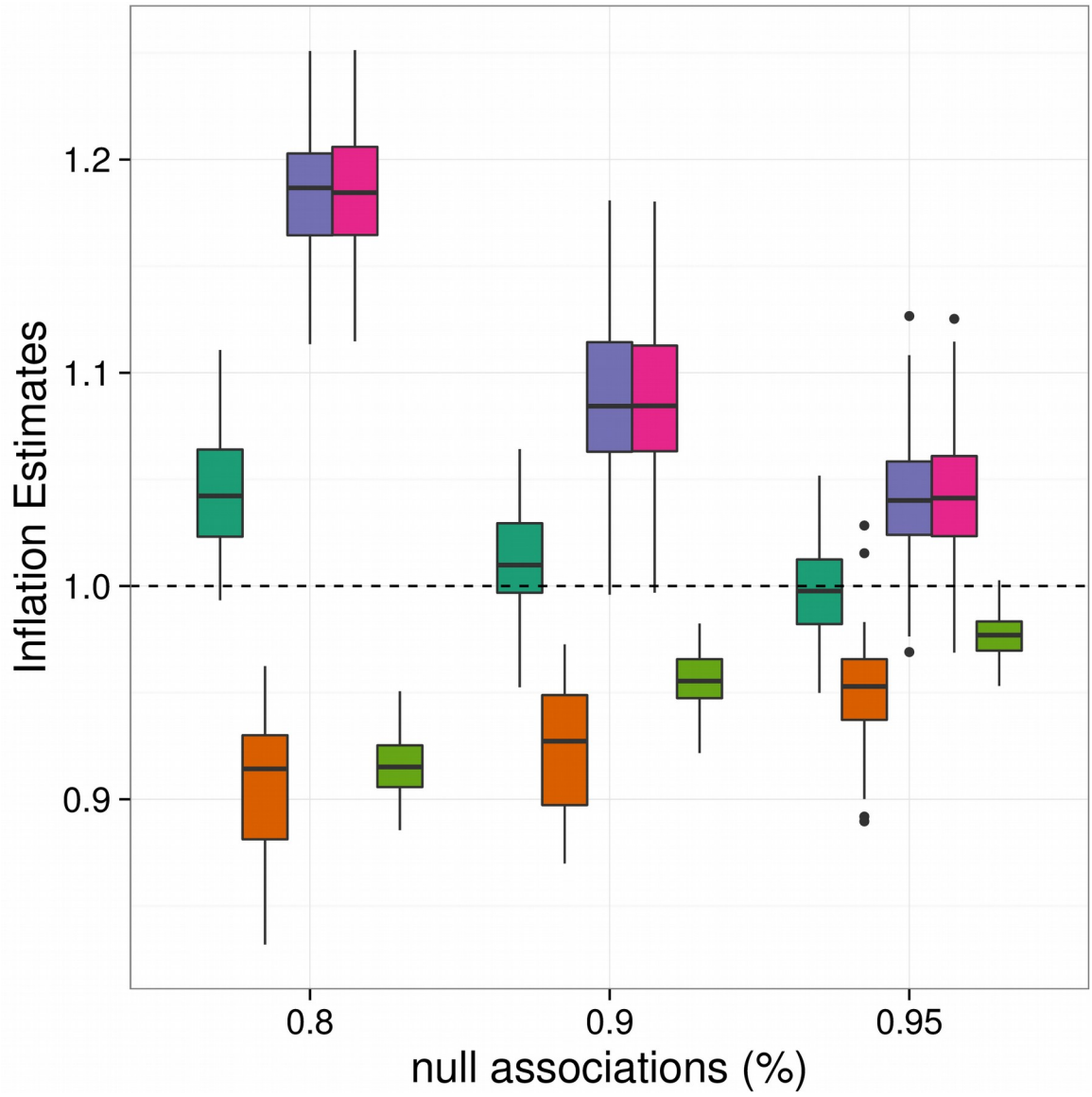

Figure 1 | Genomic inflation factor is affected by the amount of true associations
Box-plots summarizing the estimated inflation (y-axis) of 100 simulated sets of test-statistics with different amounts of true associations (x-axis). Inflation is estimated by five different methods: using our novel Bayesian approach (dark green), the central moment matching method of Efron (brown), the original genomic inflation factor (square-rooted) (purple), the median absolute deviation (mad) (dark-red) of the test-statistics and the maximum likelihood approach of Efron (light green). The original genomic inflation factor and the mad are severely biased by the amount of true associations present in the data, while our novel Bayesian inflation factor is only mildly affected.

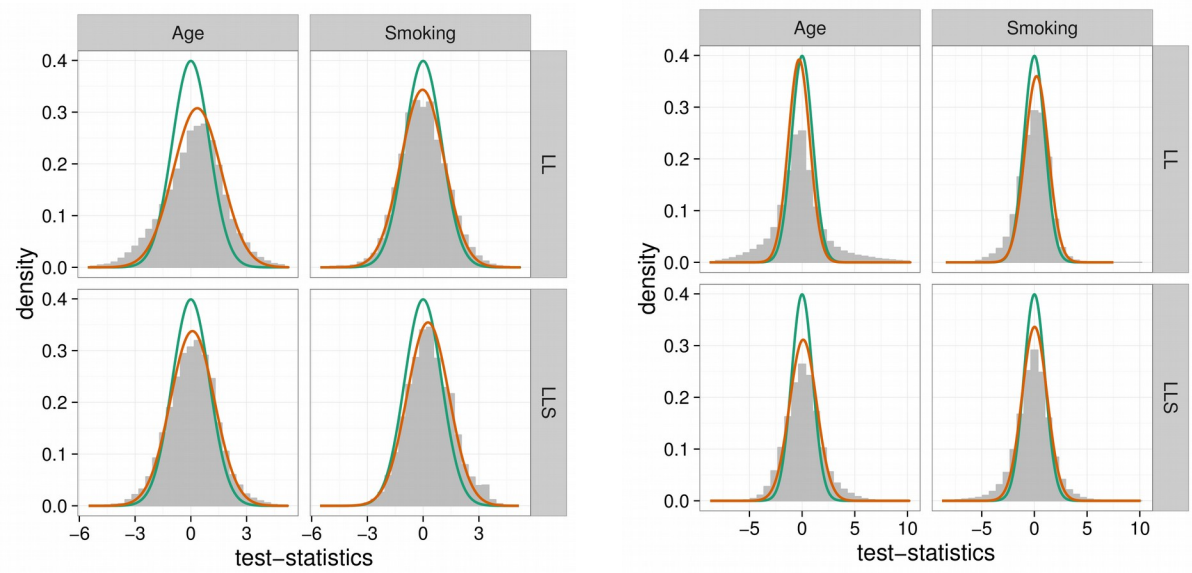

Figure 2 | Bias in epigenome- and transcriptome-wide association studies Histograms of test-statistics for TWAS (a) and EWAS (b) performed using the Leiden Longevity Study on age, and smoking status. In each panel a standard normal distribution is plotted (green) and an empirical null distribution (brown) estimated using bacon.

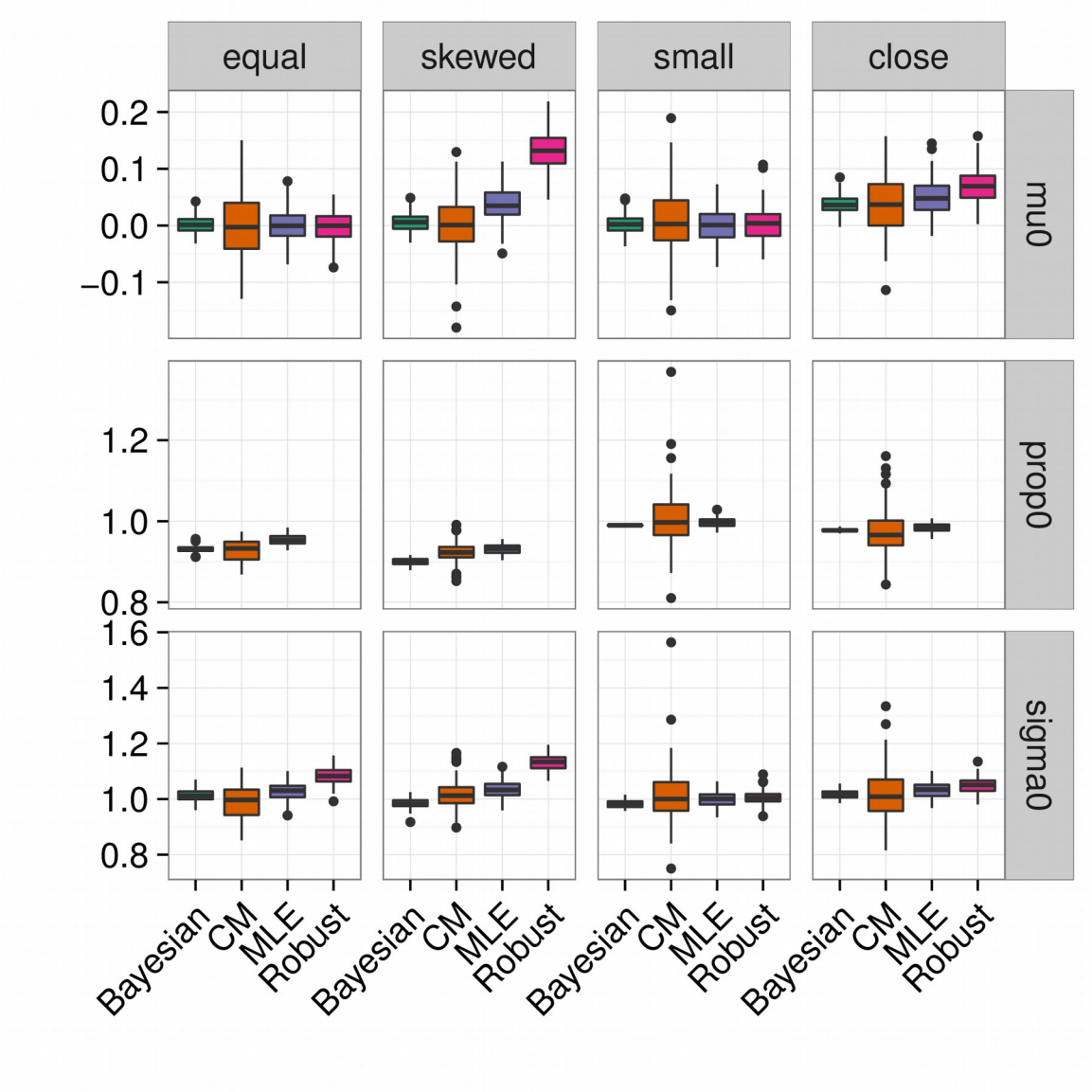

Figure 3 | Comparing different methods to estimate an empirical null distribution under different scenario's Under different scenario's (**Supplemental Methods**) the bacon inflation factor is competitive or better then existing methods that estimate an empirical null distribution, such as, central moment matching (CM)[16], maximum likelihood(MLE)[16] or robust calibration as proposed by Wang et al.[24].

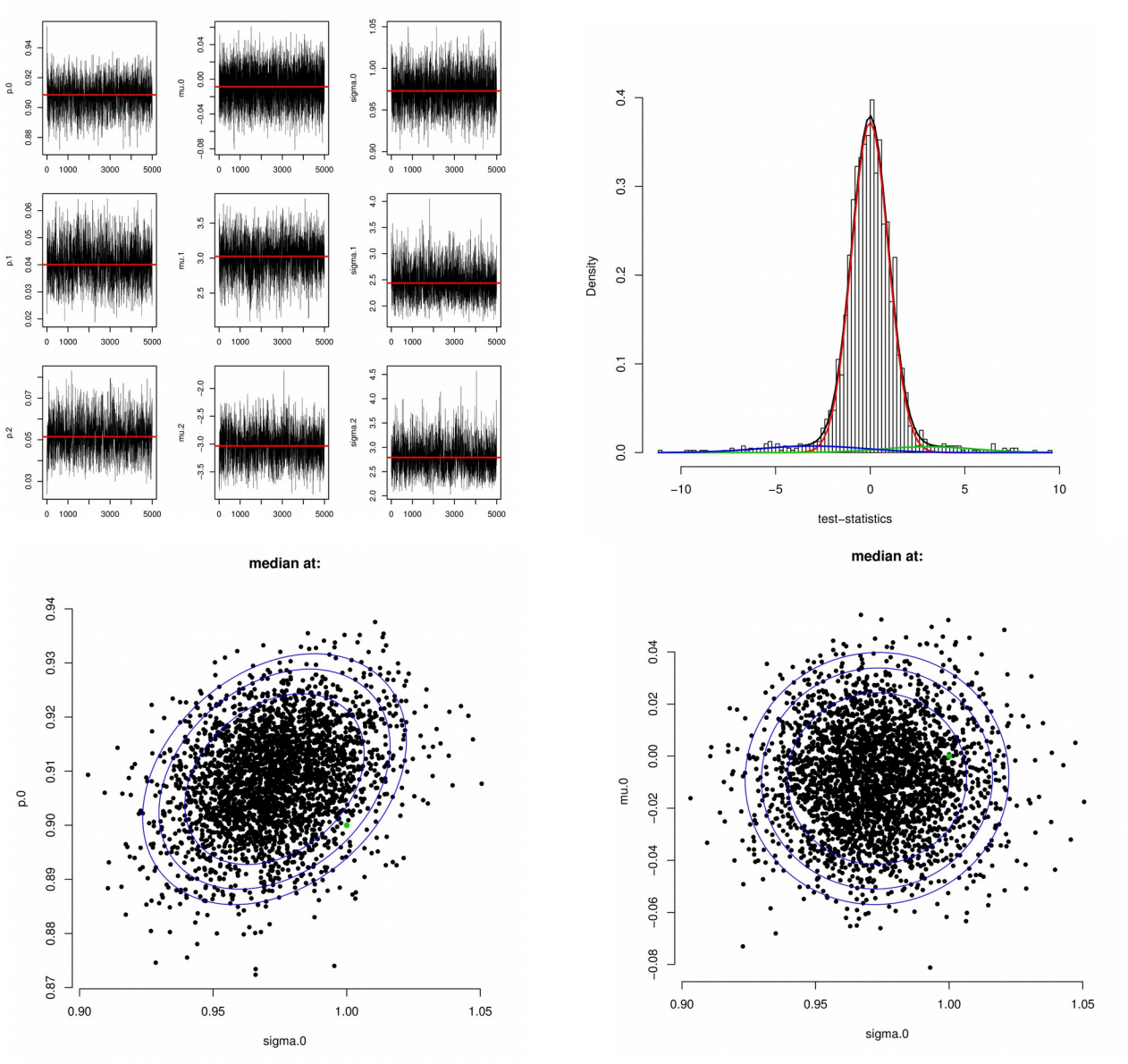

Figure 4 | bacon graphical output; diagnostic plots for the Gibbs Sampler a) Trace plot of a Gibbs Sampler run for the simulation scenario “equal”. Shown are the 2000 posterior samples of the three component normal mixture after a burn-in period of 3000 iterations. **b)** Fit of the three component normal mixture to the 2000 samples from simulation scenario “equal”. **c)** Scatter plot with normal confidence ellipses for the two, parameters proportion of null features and inflation factor. Ellipses represent from the 70, 95 and 98 percent confidence intervals. **d)** Scatter plot with normal confidence ellipses for the two, parameters proportion of null features and bias.

Supplementary Tables

Table 1 | Overview distribution demographic variables four cohorts.

	CODAM	LL	LLS	RS
Age	66 (61-71)[1]	46 (35-55)	59 (55-64)	69 (67-72)
Female sex	0.46	0.58	0.53	0.58
Smoking status	0.26, 0.59, 0.15[2]	0.47, 0.39, 0.15	0.32, 0.55, 0.13	0.35, 0.56, 0.093

[1]interquartile range
[2]fraction non-smoker, former smoker and current smoker

Table 2 | Correction for unknown batches reduces the inflation in a EWAS on age
Genomic and Bayesian inflation factors (biases) calculated from test-statistics obtained by fitting linear models with 1) only the covariate of interest 2) plus known covariates 3, 4, and 5) known covariates plus one, two and three principal component 6) plus one optimal surrogate variables estimated using iSVA[30] 7) RUVm[52] and 7) plus 3 latent variables estimated using CATE[24].

Method	Genomic inf. fac.	Bayesian inf. fac. (bias)
No	1.692	1.536 (-0.011)
Known	1.524	1.320 (0.089)
PC(1)[1]	1.390	1.245 (-0.239)
PC(2)	1.197	1.129 (0.049)
PC(3)	1.235	1.161 (0.051)
iSVA(3)	1.346	1.074 (0.057)
RUVm	1.210	1.144 (-0.082)
CATE(3)	1.327	1.191 (0.131)

[1]Within brackets the number of principal components or optimal number of surrogate variables or optimal number of latent factors.

Supplementary Tables 3 a, b, c and d as csv-files contain output of the meta-analyses, effect-sizes, standard error, P values, test-statistics of all cohorts and the meta-analysis.
These large tables will be made upon publication.