

PROCEEDINGS

Open Access



# A Bayesian mixed modeling approach for estimating heritability

Haakon E. Nustad<sup>1,2,3†</sup>, Christian M. Page<sup>4,5†</sup>, Andrew H. Reiner<sup>4</sup>, Manuela Zucknick<sup>6</sup> and Marissa LeBlanc<sup>4\*</sup>

From Genetic Analysis Workshop 20  
San Diego, CA, USA. 4 - 8 March 2017

## Abstract

**Background:** A Bayesian mixed model approach using integrated nested Laplace approximations (INLA) allows us to construct flexible models that can account for pedigree structure. Using these models, we estimate genome-wide patterns of DNA methylation heritability ( $h^2$ ), which are currently not well understood, as well as  $h^2$  of blood lipid measurements.

**Methods:** We included individuals from the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study with Infinium 450 K cytosine-phosphate-guanine (CpG) methylation and blood lipid data pre- and posttreatment with fenofibrate in families with up to three-generation pedigrees. For genome-wide patterns, we constructed 1 model per CpG with methylation as the response variable, with a random effect to model kinship, and age and gender as fixed effects.

**Results:** In total, 425,791 CpG sites pre-, but only 199,027 CpG sites posttreatment were found to have nonzero heritability. Across these CpG sites, the distributions of  $h^2$  estimates are similar in pre- and posttreatment (*pre*: median = 0.31, interquartile range [IQR] = 0.16; *post*: median = 0.34, IQR = 0.20). Blood lipid  $h^2$  estimates were similar pre- and posttreatment with overlapping 95% credibility intervals. Heritability was nonzero for treatment effect, that is, the difference between pre- and posttreatment blood lipids. Estimates for triglycerides  $h^2$  are 0.48 (*pre*), 0.42 (*post*), and 0.21 (*difference*); likewise for high-density lipoprotein cholesterol  $h^2$  the estimates are 0.61, 0.68, and 0.10.

**Conclusions:** We show that with INLA, a fully Bayesian approach to estimate DNA methylation  $h^2$  is possible on a genome-wide scale. This provides uncertainty assessment of the estimates, and allows us to perform model selection via deviance information criterion (DIC) to identify CpGs with strong evidence for nonzero heritability.

## Background

Narrow-sense heritability ( $h^2$ ), traditionally estimated using twins or other constrained family relationships, can also be estimated in wider pedigrees using a linear mixed-model approach [1, 2]. This approach is well-established for traits having moderate to high  $h^2$  [2, 3]. What is less clear is how well these models perform for traits with low  $h^2$ , which is the case for some proportion of cytosine-phosphate-guanine (CpG)

methylation sites genome wide. This is of particular interest for  $h^2$  because it is a ratio of variance components, that is, the proportion of phenotypic variance explained by additive genetic variance. Compared to estimates of the mean, estimates of variance tend to have large uncertainty.

DNA methylation (DNAm) is an epigenetic mark that is implicated in many heritable diseases and traits. DNAm patterns are influenced by the environment, change over lifetime, and exhibit mitotic heritability [4]. There are multiple reports of DNAm  $h^2$  with CpG-specific estimates ranging from 0 to 1, and with genome-wide mean or median estimates ranging from approximately 15 to 30% [5–7].

\* Correspondence: [marissa.leblanc@medisin.uio.no](mailto:marissa.leblanc@medisin.uio.no)

†Haakon E. Nustad and Christian M. Page contributed equally to this work.

<sup>4</sup>Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Klaus Torgårds vei 3, 0372 Oslo, Norway

Full list of author information is available at the end of the article



Any point estimates of  $h^2$  should ideally be accompanied by an assessment of the uncertainty. A Bayesian mixed model with inference by integrated nested Laplace approximations (INLA) [2, 8] presents an improvement on the traditionally used simple, balanced designs (eg, parent-offspring or twin-based correlations) as it can include the full pedigree structure. Additionally, full posterior inference provides uncertainty for the estimated  $h^2$ , which is not obtainable using frequentist methods, and allows an assessment of whether the true  $h^2$  is nonzero via comparison of models with and without a genetic component.

In this article, we use three-generation pedigrees from the Genetics of Lipid Lowering Drug and Diet Network (GOLDN) [9] to estimate  $h^2$  in a fully Bayesian framework. Blood lipid and DNAm measurements were available pre- and posttreatment with a lipid-lowering drug (fenofibrate). We first investigate continuous blood lipid traits with known moderate to high  $h^2$  and compare the Bayesian estimates with previously published frequentist estimates [10]. Second, we estimate the genome-wide distribution of DNAm  $h^2$ .

The Bayesian mixed-model approach provides posterior distributions and therefore uncertainty for  $h^2$ , the additive genetic variance, and the environmental variance. We further investigate if it is reasonable to estimate  $h^2$  for each CpG site given the data by performing model selection, to assess if the genetic component should be included at all. When the additive genetic variance is truly zero, a linear mixed model including the family structure will misspecify the model, making any resultant  $h^2$  estimates untrustworthy. When  $h^2$  is truly nonzero but smaller than the minimum  $h^2$  detectable in a given data set (ie, when the study is underpowered), we would expect extremely large standard errors for the  $h^2$  estimates.

In this article, our aims are to estimate  $h^2$  for (a) blood lipid traits and (b) genome-wide DNAm using a Bayesian approach, including a model selection step to identify CpG sites with strong evidence for nonzero heritability. We highlight the importance of the model selection step and of quantifying the uncertainty in  $h^2$  estimates.

**Methods**

**Data**

Individuals in the GOLDN study having DNAm and blood lipid measurements, as well as information on gender and age, were included in the analysis [11]. Genome-wide DNAm was measured using the 450 K Infinium array at 463,995 CpG sites and was available for 995 and 530 individuals at pre- and postfenofibrate treatment, respectively. From the original 463,995 sites available, we removed a list of known single nucleotide polymorphism-related CpG sites. Blood lipid levels were available both pre- and post-treatment for 818 and 861 individuals for triglycerides (TGs) and high-density lipoproteins (HDLs), respectively.

In our analyses, we use log-transformed TG and HDL (averaged over 2 measurements taken 1 day apart), and simply refer to these transformed average values as TG and HDL from here on.

Standard quality control inspection of the DNAm data showed systematic differences with respect to probe-type chemistry. We therefore normalized the data with respect to the probe type using beta-mixture quantile normalization using the R package *wateRmelon* [12]. We calculated the M values for each probe and used these in the analysis.

**Statistical model**

We estimate  $h^2$  in a Bayesian framework using the INLA package in Rue et al. [8]. Using the following model, we investigate  $h^2$  for TG, HDL, and DNAm in the GOLDN study pre- and posttreatment:

$$\begin{aligned}
 y_i \mid \eta_i, \sigma_e^2 &\sim N(\eta_i, \sigma_e^2) \forall i = 1, \dots, n \text{ with } \eta_i = \\
 &\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + u_i \\
 \mathbf{u} \mid \sigma_g^2 &\sim N(\mathbf{0}, 2\sigma_g^2 \mathbf{K}) \text{ with } \mathbf{u} = (u_1, \dots, u_n)
 \end{aligned}
 \tag{1}$$

where  $\mathbf{y} = (y_1, \dots, y_n)$  is the outcome vector, that is, DNAm for a given CpG, TG, or HDL, for  $n$  subjects.  $N(\eta_i, \sigma_e^2)$  is the likelihood function for subject  $i$ , where  $\eta_i$  is the linear predictor defining the latent variables of interest and  $\sigma_e^2$  denotes the residual unexplained variance in the model. Fixed effects are the intercept  $\beta_0$  and  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  corresponding to the effects of age ( $x_{1i}$ ) and gender ( $x_{2i}$ ). Family structure was modeled with  $\mathbf{u}$  as a random effect, with distribution  $N(\mathbf{0}, 2\sigma_g^2 \mathbf{K})$ , where  $\mathbf{K}$  is the kinship matrix calculated based on pedigree information using the *{kinship2}* package in R [13], and  $\sigma_g^2$  is the additive genetic variance. We assume default noninformative priors for all hyperparameters for the genome-wide investigation of  $h^2$  (in particular  $\sigma_*^{-1} \sim \text{Gamma}(1, 0.00005)$ ), but other priors were also investigated. From this model,  $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$ .

With the deviance information criterion (DIC) [14, 15], we can investigate the fit of a model to the data and do model selection. We can compare the DIC of the 2 latent models:

$$\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}
 \tag{2}$$

$$\eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + u_i
 \tag{3}$$

and use the recommended  $\Delta \text{DIC} = \text{DIC}_{(2)} - \text{DIC}_{(3)} > 10$  to choose if we should include the genetic component ( $u_i$ ) or not.

As a result of the setup of INLA, we can only obtain point estimates of  $h^2$  with this model formulation, but not the marginal posterior distribution. If this is of interest, the following equivalent model formulation must be

used instead [2]. The nuisance parameter  $\sigma_0^2$  is introduced to ensure that  $\sigma_e^2$  remains identifiable. It is fixed at a small value, so that it does not absorb much of the variation in the data and thus does not interfere with  $\sigma_g^2$  or  $\sigma_e^2$ . We used  $\sigma_0^2 = 4.54e-5$ , which corresponds to  $\log(\sigma_0^2) = -10$  and was suggested in an earlier publication [2].

$$y_i \mid \eta_i, \sigma_e^2 \sim N(\eta_i, \sigma_e^2) \forall i = 1, \dots, n \text{ with } \eta_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + u_i + \varepsilon_i$$

$$\mathbf{u} \mid \sigma_g^2 \sim N(\mathbf{0}, 2\sigma_g^2 \mathbf{K}) \text{ with } \mathbf{u} = (u_1, \dots, u_n)$$

$$\boldsymbol{\varepsilon} \mid \sigma_e^2 \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}) \text{ with } \boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$$

**Results**

Figure 1 shows the marginal posterior distribution of  $h^2$  for HDL and TG levels pre- and posttreatment, as well as for the change (post-pre). The change in blood lipid levels can be interpreted as the response to treatment. The model selection step supported the inclusion of  $u_i$  in the model and thus estimation of  $h^2$  for each case. The mode estimates of  $h^2$  for TG are 0.48, 0.42, and 0.21 for baseline, posttreatment, and change, respectively. Likewise, the mode estimates for  $h^2$  for HDL are 0.61, 0.68, and 0.10. Based on point estimates alone, HDL appears to be more heritable than TG but their 95% credible intervals overlap. TG is more heritable pre-treatment than posttreatment, but the opposite is true for HDL. However, the 95% credibility intervals for  $h^2$  pre- and posttreatment overlap each other substantially for both traits.

In Fig. 2, the histogram of point estimates of  $h^2$  of DNAm for 448,040 CpG sites is displayed, for both pre- and posttreatment. Model selection indicates if there is evidence for nonzero  $h^2$ .

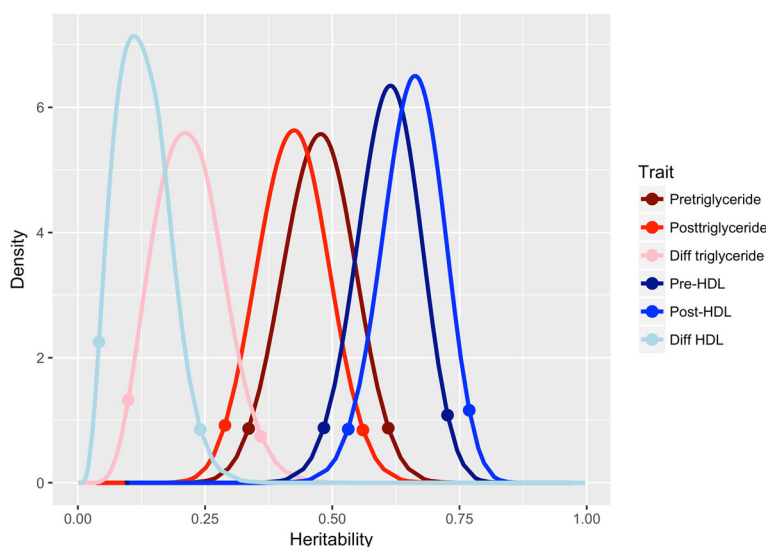
For pretreatment, we detect nonzero  $h^2$  for 425,791 of 448,040 CpG sites (95.0%). For the CpG sites with strong evidence for nonzero  $h^2$ , the mean, median, and inter-quartile range (IQR) are 0.33, 0.31, and 0.16. In contrast, for posttreatment we detect nonzero  $h^2$  for 199,027 of 448,040 CpG sites (44.4%) with mean, median, and IQR equal to 0.36, 0.34, and 0.20. Both cases show evidence of zero inflation and a right skewed distribution of DNAm  $h^2$ . Indirectly, we see that GOLDN is underpowered for detecting  $h^2$  of DNAm less than approximately 10 to 15%.

Figure 2 clearly shows a striking difference in the distribution of genome-wide  $h^2$  estimates between pre- and post-treatment. Figure 3 shows the direct pairwise comparison of the two  $h^2$  estimates for DNAm across those CpG sites with strong evidence for nonzero  $h^2$ . For these CpG sites, we observe a moderate Pearson correlation between the baseline and posttreatment heritability estimates of 0.53.

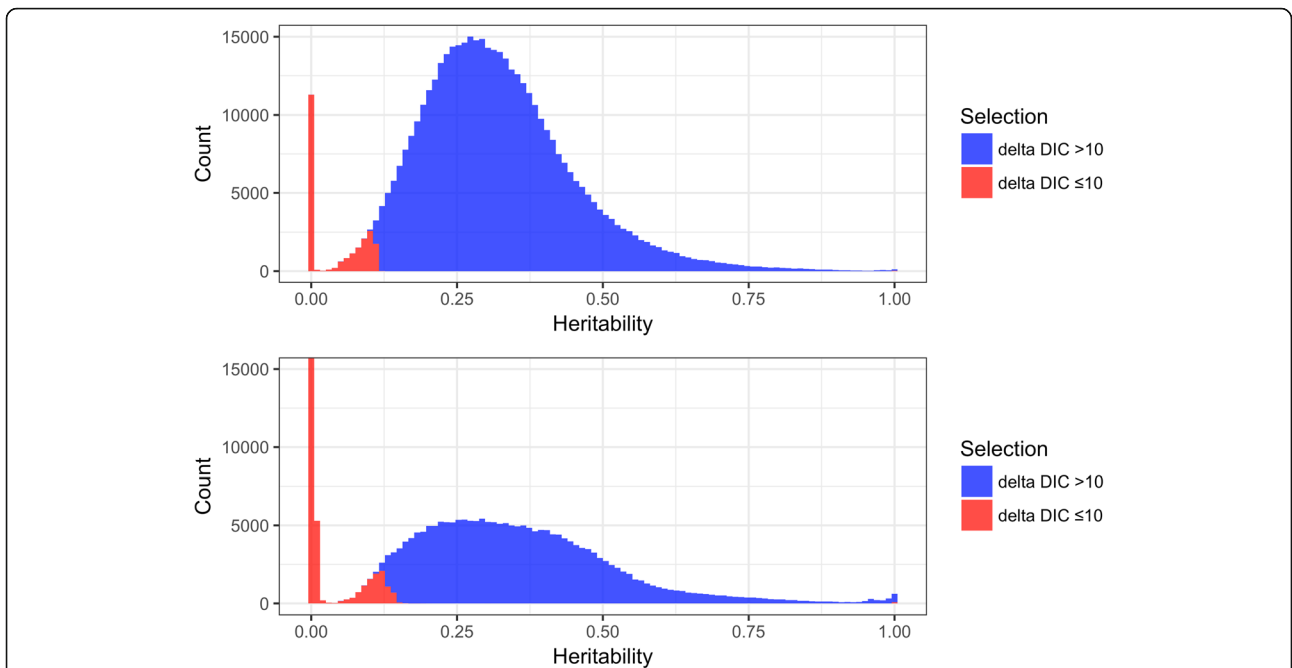
**Discussion and conclusions**

In this article, we use a Bayesian approach to estimate  $h^2$  for blood lipid levels and DNAm in extended pedigrees, and show that it is feasible to do genome-wide Bayesian estimation of  $h^2$ .

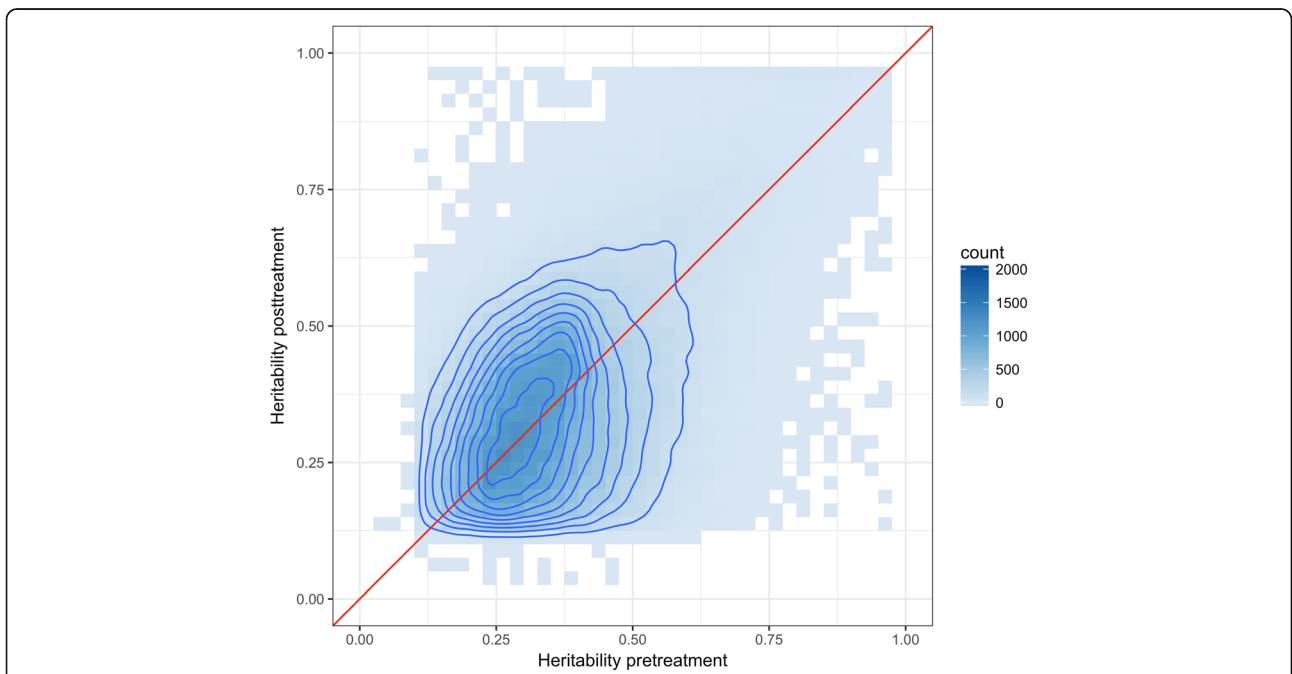
For blood lipids, our pretreatment point estimates for  $h^2$  using the Bayesian approach ( $h^2_{TG} = 0.48$  and  $h^2_{HDL} = 0.61$ ) are comparable to previously reported estimates using other methods [16]. Notably, the 95% credibility intervals are wide (see Fig. 1). Interestingly, the response to treatment (change in blood lipid levels from



**Fig. 1** Heritability ( $h^2$ ) estimates for high-density lipoprotein (HDL) and triglycerides (TG). Posterior  $h^2$  distribution for HDL and TG pre- and posttreatment, as well as for the change from pre- to posttreatment ("Diff"). The points indicate the 95% credibility region for each  $h^2$  estimate



**Fig. 2** Genome-wide heritability ( $h^2$ ) estimates for DNA CpG methylation (DNAm). Histogram of point estimates for DNAm  $h^2$  genome wide. The top panel shows the  $h^2$  estimates pretreatment and the lower panel shows the  $h^2$  estimates posttreatment. The blue color is for CpG sites with strong evidence for heritability (“delta DIC >10”) and the red color is for the CpG sites with no evidence for heritability given the data (“delta DIC  $\leq$ 10”), based on model selection comparing the models with and without pedigree information [eqs. (2) and (3)]. The two colors are stacked on top, so that both colors are visible on each bar. The actual height of the zero-stack bar for the posttreatment is 238,165 CpG sites



**Fig. 3** Contour plot of correlation of DNA CpG methylation heritability ( $h^2$ ) pre- and posttreatment. The plot is based on the 194,741 CpG sites that have evidence for nonzero heritability in both pre- and posttreatment methylation measurements

pre- to posttreatment; see Fig. 1) also appears to be heritable. This may have important implications for personalized medicine, especially if one were able to identify explanatory variables (genetic or environmental) that predict the individual genetic value ( $u_i$ ).

The genome-wide distribution of DNAm  $h^2$  (as ascertained by the Illumina Infinium 450 K chip) appears to follow a 2-group mixture model with a  $h^2 = 0$  component and a nonzero component following a smooth, unimodal right skewed distribution. The mixture proportion is strikingly different in the pre- and posttreatment (see Fig. 2). It is unlikely that treatment with fenofibrate could cause substantial genome-wide changes in DNAm  $h^2$ . It is much more likely that technical differences between the DNAm time points explain the difference [17].

The  $h^2$  estimates we present in this article are generally similar to earlier genome-wide results in different populations [5, 7]. These previously published studies also show zero inflation, although the reported proportions vary substantially, from 5 to 17%. All show right-skewed distributions, but with slightly different modality. We report the median,  $h^2_{\text{pre}} = 0.31$ ,  $h^2_{\text{post}} = 0.34$ , in the GOLDN study only for CpG sites with strong evidence for nonzero  $h^2$  (ie,  $\Delta\text{DIC} > 10$ ). This is higher than previously reported, although this is to be expected if previous reports included the zero component in the calculation of the median. With the zero component included, our median estimates decrease to  $h^2_{\text{pre}} = 0.30$ ,  $h^2_{\text{post}} = 0.0036$ .

A Bayesian model holds some advantages over a frequentist model when estimating  $h^2$ . The most notable difference is that a Bayesian model can estimate the full posterior distribution of the parameters in question. This includes the random effect variance parameters, which allows more direct assessment of the uncertainty of the random-effect estimates—and thus of the  $h^2$  estimates—than could be achieved with classical mixed-effects models. For example, by comparing posterior credibility intervals for  $h^2$  pre- versus posttreatment, we can determine if the  $h^2$  changes are a result of treatment (see Fig. 1). In addition, model selection can be performed in a principled way by DIC or Bayes factors to assess if the data provide sufficient evidence for nonzero  $h^2$ . When the  $h^2$  is truly zero, the genetic values  $u_i$  should not be included in the model, and therefore a model without the genetic values should be favored. Indeed, DIC-based model selection favored the model without random effect in the vast proportion of CpG sites for which  $h^2$  was estimated to be close to zero.

An important limitation in our study is the lack of information on shared households to estimate the effect of shared environment. This likely results in an overestimation of  $h^2$ . The CpG sites on the 450 K array are a nonrandom selection of CpG sites with a bias toward CpG islands that often have regulatory function. This is important to consider, as an oversampling of biologically

important CpG sites could result in a biased estimation for nonspecific genome-wide DNAm  $h^2$ .

In this article, we use a Bayesian approach to estimate  $h^2$  in extended pedigrees and show the importance of using model selection to determine if there is strong evidence of nonzero  $h^2$  given the data. Future work should focus on simulation studies comparing the frequentist and Bayesian approaches, with particular focus on traits having low  $h^2$ .

#### Acknowledgements

Part of this work was performed on the Abel Cluster, owned by the University of Oslo. The authors also acknowledge input from Robert Lyle and Kristina Gervin.

#### Funding

Publication of this article was supported by NIH R01 GM031575. HEN received funding from the Norwegian Research Council. CMP and ML received internal funding from Oslo University Hospital.

#### Availability of data and materials

The data that support the findings of this study are available from the Genetic Analysis Workshop (GAW) but restrictions apply to the availability of these data, which were used under license for the current study. Qualified researchers may request these data directly from GAW.

#### About this supplement

This article has been published as part of BMC Proceedings Volume 12 Supplement 9, 2018: Genetic Analysis Workshop 20: envisioning the future of statistical genetics by exploring methods for epigenetic and pharmacogenomic data. The full contents of the supplement are available online at <https://bmcpoc.biomedcentral.com/articles/supplements/volume-12-supplement-9>.

#### Authors' contributions

HEN, ML, MZ, and CMP conceived and designed the experiment; HEN, ML, AHR, and CMP generated the results; and HEN, ML, MZ, and CMP interpreted the results. All authors wrote the manuscript and approved the final version.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Medical Genetics, Oslo University Hospital, Kirkeveien 166, 0450 Oslo, Norway. <sup>2</sup>Department of Medical Genetics, University of Oslo, Klaus Torgårds vei 3, 0372 Oslo, Norway. <sup>3</sup>PharmaTox Strategic Research Initiative, University of Oslo, Sem Sælands vei 3, 0371 Oslo, Norway. <sup>4</sup>Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Klaus Torgårds vei 3, 0372 Oslo, Norway. <sup>5</sup>Department of Non-Communicable Disease, Norwegian Institute of Public Health, Marcus Thranes gate 6, 0473 Oslo, Norway. <sup>6</sup>Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Sognsvannsveien 9, 0372 Oslo, Norway.

Published: 17 September 2018

#### References

1. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet.* 1998;62(5):1198–211.



2. Holand AM, Steinsland I, Martino S, Jensen H. Animal models and integrated nested Laplace approximations. *G3 (Bethesda)*. 2013;3(8):1241–51.
3. Ralston SH, Uitterlinden AG. Genetics of osteoporosis. *Endocr Rev*. 2010; 31(5):629–62.
4. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics*. 2009;182(3):845–50.
5. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, Painter JN, Martin NG, Visscher PM, Montgomery GW. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome Biol*. 2014; 15(5):R73.
6. Quon G, Lippert C, Heckerman D, Listgarten J. Patterns of methylation heritability in a genome-wide analysis of four brain regions. *Nucleic Acids Res*. 2013;41(4):2095–104.
7. Kulkarni H, Kos MZ, Neary J, Dyer TD, Kent JW Jr, Goring HH, Cole SA, Comuzzie AG, Almasy L, Mahaney MC, et al. Novel epigenetic determinants of type 2 diabetes in Mexican-American families. *Hum Mol Genet*. 2015; 24(18):5330–44.
8. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J R Stat Soc Series B Stat Methodol*. 2009;71:319–92.
9. Irvin MR, Kabagambe EK, Tiwari HK, Parnell LD, Straka RJ, Tsai M, Ordovas JM, Arnett DK. Apolipoprotein E polymorphisms and postprandial triglyceridemia before and after fenofibrate treatment in the genetics of lipid lowering and diet network (GOLDN) study. *Circ Cardiovasc Genet*. 2010;3(5):462–7.
10. Weiss LA, Pan L, Abney M, Ober C. The sex-specific genetic architecture of quantitative traits in humans. *Nat Genet*. 2006;38(2):218–22.
11. Das M, Irvin MR, Sha J, Aslibekyan S, Hidalgo B, Perry RT, Zhi D, Tiwari HK, Absher D, Ordovas JM, et al. Lipid changes due to fenofibrate treatment are not associated with changes in DNA methylation patterns in the GOLDN study. *Front Genet*. 2015;6:304.
12. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14:293.
13. Therneau T, Atkinson E, Sinnwell J, Matsumoto M, Schaid D, McDonnell S: Package “kinship2” [Online] 2012. Available at: <http://cran.r-project.org/web/packages/kinship2/kinship2.pdf>.
14. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Series B Stat Methodol*. 2002;64:583–616.
15. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*, vol. 2. Boca Raton: Chapman & Hall/CRC; 2014.
16. Kathiresan S, Manning AK, Demissie S, D’Agostino RB, Surti A, Guiducci C, Gianniny L, Burt NP, Melander O, Orho-Melander M, et al. A genome-wide association study for blood lipid phenotypes in the Framingham heart study. *BMC Med Genet*. 2007;8(Suppl 1):S17.
17. LeBlanc M, Nustad HE, Zucknick M, Page CM. Quality control for Illumina 450K methylation data in the absence of iDat files using correlation structure in pedigrees and repeated measures. *BMC Genet*. 2018;19(Suppl 1) <https://doi.org/10.1186/s12863-018-0636-5>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

